I hereby certify that this paper (along with any paper referred to as being attached or enclosed) is being transmitted via the Office electronic filing system in accordance with § 1.6(a)(4).

Dated: July 7, 2008 Signature: /David A. Gass #38,153/ (David A. Gass)

Docket No.: 30847/2048-004
(PATENT)

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Patent Application of:
Anna Helgadottir

Application No.: 10/829,674 Confirmation No.: 6838

Filed: April 22, 2004 Art Unit: 1634

For: SUSCEPTIBILITY GENE FOR MYOCARDIAL INFARCTION AND STROKE Examiner: J. A. Goldberg

DECLARATION OF GUDMAR THORLEIFSSON UNDER 37 C.F.R. § 1.132

Docket No.: 30847/2048-004

(PATENT)

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Patent Application of:
Anna Helgadottir

Application No.: 10/829,674

Confirmation No.: 6838

Art Unit: 1634

Filed: April 22, 2004

Examiner: J.A. Goldberg

For: SUSCEPTIBILITY GENE FOR MYOCARDIAL INFARCTION AND STROKE

DECLARATION OF GUDMAR THORLEIFSSON UNDER 37 C.F.R. § 1.132

1. I am a scientist and statistician at deCODE genetics, ehf. ("deCODE"), having worked at deCODE for 9 years. My *curriculum vitae* is attached hereto.

I. INTRODUCTION

2. I have been Group Leader for the Cardiovascular Disease Group in the Statistical Department at deCODE since 2001. I was involved in the the research team at deCODE that identified at risk-variants for myorcardial infraction (MI) within the gene encoding 5-lipoxygenase activating protein (FLAP). I am providing this declaration to make available to the United States Patent and Trademark Office (PTO) a further explanation of the data supporting the association of variants within the FLAP gene with with increaased risk of Myocardial Infarction.

II. OBSERVATION AND ANALYSIS REGARDING THE RELIABILITY OF GENETIC ASSOCIATION STUDIES

3. I have read the PTO's examination reports dated July 21, 2006; April 16, 2007; and January 4, 2008, prepared by Examiner Goldberg. The information and opinions that I provide in this section pertain to the PTO's positions in the examination reports about the unpredictability and alleged irreproducibility of genetic association studies.

4. Historically, the field of human genetics has been plagued by claims of association findings, which have on scrutiny not held up in replication studies. There

are many reasons for this lack of replication, including poor experimental design, population stratification and lack of power in the initial discovery and/or replication studies. However, important advances have been made in the last few years, and in particular in the last couple of years. With the discovery and validation of very large numbers of Single Nucleotide Polymorphisms (SNPs), a resource for genome-wide searches for SNPs associated with human disease has been generated (see http://www.hapmap.org). Technological advances have also made highly accurate genotyping of up to one million SNPs on a single chip possible. Finally, experimental design in genetic association studies has been drastically improved, and there is a general consensus requirement in the scientific community that all initial findings should be replicated (i.e., confirmed) prior to publication.

5. I would like to clarify the concept of "statistical power" in genetic association studies, as this is important to understand why some studies are less likely than others to replicate an observed genetic association, even though the underlying association is real and applies to all populations. Power is simply the probability of finding a statistically significant difference between cases and controls in a particular study, given a real underlying difference between cases and controls. The power of a genetic association study depends on several factors, most important being the size of the study populations (*e.g.*, cohort size; numbers of cases and controls), but also the effect size or the odds ratio (OR) of the genetic variant and the frequency of the variants. "Odds ratio" is a measure of effect size; it can be defined as the ratio of the odds of an event occurring in one group to the odds of it occurring in another group. An odds ratio of 1 implies that the event is equally likely in both groups. When an odds ratio is greater than 1 when comparing cases to controls, it implies that a particular variant is more likely to be found in the case group than in the controls. The power of an association study increases both with increased cohort size and larger OR of the variant. Hence, for variants with a small effect, much larger study groups are needed to detect (with statistical significance) the association than for a variant with a large effect. As an example, given the estimated effect of HapA based on the replication groups, OR = 1.11, and assuming that the population frequency of HapA is 14%, a study group of 1,000 cases and 1,000 controls only has about 21% chance of detecting a significant difference in the frequency of the HapA haplotype between MI cases and controls (for the power calculation method, see *e.g.* Rice JA, Mathematical

Statistics and Data Analysis, 2nd ed, Duxbury Press, Belmond, California (1995)). If the study group is increased to 5,000 cases and 5,000 controls, the power is increased to a 74% chance of detecting significant association to HapA.

6. A large number of studies meeting established criteria of statistical significance have been published in the last 2 years, many of which are multicenter Genome-Wide Association (GWA) studies (reviewed in Pearson TA & Manolio, TA, JAMA 299:1335-1344 (2008); Bowcock, AM, Nature 447:645-46 (2007); Altshuler, D & Daly, M., Nature Genet 39:813-815 (2007); Kruglyak, L, Nature Reviews Genet 9:314-18 (2008); Kingsmore, S.F., et al., Nature Reviews Genet 7:221-230 (2008); Frayling, TM, Nature Reviews Genet 8:657-662 (2007)). Through GWA studies, nearly 100 loci for up to 40 common diseases have been identified and confirmed (Pearson TA & Manolio, TA, JAMA 299:1335-1344 (2008)).

7. A common theme for most of the risk-associated genetic variants thus identified is that the risk conferred by each variant is typically relatively small, e.g., usually a relative risk in the range of about 1.1 – 1.5 (Bowcock, p. 646; Pearson, p. 1341, Altshuler, p. 814; Frayling, p.659)). For some diseases, a large number of variants have been identified. A good example is Type 2 diabetes, which was previously referred to as the "geneticist's nightmare" due to the lack of identification of genetic variants for the disease. As summarized in Frayling and in Zeggini et al (Nature Genetics Mar 30 [Epub ahead of print] (2008)), a total of 17 genomic regions that alter the risk of Type 2 dibabetes have been identified and solidly validated. This has been possible by combining results from multiple centers individually investigating the genetics of Type 2 diabetes in a joint meta-analysis of available data (see Zeggini et al). What is even more remarkable is that with the exception of the common variant in the TCF7L2 gene originally identified by scientists at deCODE genetics, the risk for these genetic variants is quite small, ranging from 1.09 to 1.20 per copy of the variant (the risk conferred by the TCF7L2 variant is about 1.35 per copy).

8. The plethora of variants for the common diseases identified in the last couple of years has thus established that genetic risk for most common diseases appears to be modulated through multiple low-risk variants which, in any given individual, act in a concerted manner, together with environmental risk factors, to

establish overall predisposition for the particular disease. Thus, scientists in the field are accepting the validity of genetic association studies that demonstrate low, but statistically significant risk in replication studies.

9. The Patent Office has cited an article by Ionnidis that allegedly states, "As a general rule of thumb we are looking for a relative risk of three or more [before accepting a paper for publication], particularly if it is biologically implausible or if it's a brand new finding." (See, e.g., Office action dated January 4, 2008, at p. 11.) The papers that I cite above – from well-respected journals – demonstrate that relative risk of three or more is NOT an accepted threshold criteria in the field for publication. Nor does it reflect the current thinking or data involving the effects of genetic variation and common diseases.

III. THE RELATIVE RISK SCORES IN THE PRESENT APPLICATION ARE CONSISTENT WITH RISKS REPORTED IN THE LITERATURE AND GENERALLY ACCEPTED AS "REAL" RISK SCORES IN THE FIELD.

10. The method used by Dr. Manolescu in his meta-analysis of available data, and by Dr. Helgadottir in her meta-analysis, represents standard methodology commonly used to combine results from multiple genetic association studies. The Mantel-Haenszel model that they used (Mantel and Haenszel, 1959; Woodward, 2005) is designed to deal with the situation where association results from different populations, with possible different population frequency of the genetic variant, are combined. In that case, the model combines the results assuming that the effect of the variant on the risk of the disease, as measured by the OR, is the same in all populations, while the frequency of the variant may differ between the populations.

11. I am familiar with the contents of the declarations provided by Anna Helgadottir (executed 22 January 2007) and Andrei Manolescu (executed 16 October 2007) previously officially filed with the PTO, as well as the Table informally provided for interview purposes, before the Manolescu declaration was completed.

12. The Tables provided in the Declarations (and interview) describe results of "meta-analysis" of association data for a four-marker FLAP haplotype (called "HapA") with MI. Meta-analyses are commonly used as a quantitative way of combining the results of several genetic association studies, so as to provide an

overall estimate of the underlying genetic effect on phenotype. Such analyses are especially valuable to achieve acceptable statistical power when genetic effects on phenotype are small (e.g., in the range of increased relative risk of 1.1 or 1.2), because individual studies may be insufficiently powered (due to limited sample size) to detect the true genetic effect. Even in the absence of other factors such as poor study design, population stratification and phenotype heterogeneity, the results of individual association studies are expected to fluctuate due to inherent fluctuations in sampling: If one examines an ensemble of studies, such fluctuations due to sampling are larger with respect to smaller and less powered study groups. One way for designers of studies to address the statistical reality of such sampling fluctuations is to use sufficiently large cohorts for the association study. Alternatively, it is possible to achieve a large effective sample size by combining results from several small studies. However, if the study cohorts that are combined come from different populations, then the population frequencies of the genetic variant tested may differ between the studies, and appropriate (statistically accepted) methods commonly used in meta-analysis must be applied when combining results from different studies. We at deCODE used such statistical tools to perform the meta-analysis presented in the declarations of Dr. Helgadottir and Dr. Manolescu, as I explained above in the preceding section of this declaration. It is important to note that, while individual studies may not yield significant results, the combined results in such meta-analyses can be very significant if there is consistency in the direction of the observed effect across the studies included in the analysis.

IV. EXPLANATION OF DIFFERENCES IN THE DATA SETS AND STATISTICAL ANALYSES IN THE HEGADOTTIR AND MANOLESCU DECLARATIONS.

13. The Examiner notes several differences in the Tables provided in the declarations of Dr. Helgadottir, Dr. Manolescu, and in the interview Table provided July 31, 2007. The first thing to note is that the interview Table from July 31, 2007, reports two-sided P-values, while the two Declarations report one-sided P-values. Correcting for this, the interview Table agrees with the Table provided in the declaration of Dr. Manolescu. The choice of two-sided P-values in the interview Table is unnecessarily conservative, since the analysis is testing a specific effect with

known direction (i.e., increased risk). In such circumstances, it is statistically appropriate to report a one-sided P-value.

14. The differences between the analysis presented in the declarations of Dr. Helgadottir and Dr. Manolescu, that involve the results for the case/control groups from Philadelphia, Cleveland, Atlanta and Durham, arise primarily due to the fact that the analyses presented in the two declarations were performed by deCODE at different time points, using in each case the most recent phenotype information and genotype data available at that time.

15. More specifically, for the Philadelphia cohort the number of controls, and the calculated P-values differed in the declarations because, subsequent to the analysis of Dr. Helgadottir, updated phenotype information revealed that three of the controls had coronary artery disease. Hence, these three controls in the dataset used by Helgadottir were excluded from the analysis of Dr. Manolescu, as individuals with known coronary artery and related diseases are excluded from all the control groups. Although this change had a very small effect on the estimated frequency of the haplotype in controls, the change led to a small change in the P-value presented in the analysis of Dr. Helgadottir and Dr. Manolescu, respectively.

16. For the Cleveland cohort, the differences between the two analyses also are due to updated phenotype information. In the updated information, five additional individuals were reported with MI and hence included in the "cases" group in the analysis of Dr. Manolescu. Three of these individuals with MI had been included as controls in the analysis presented in the declaration of Dr. Helgadottir. Additional phenotype information also showed that several of the controls used in the analysis presented in the declaration of Dr. Helgadottir had been diagnosed with either abdominal aortic aneurysm or peripheral artery disease, both of which are considered atherosclerotic diseases. Hence, those individuals were excluded from the analysis of Dr. Manolescu.

17. For the Atlanta cohort, the differences between the two declarations result from both updated phenotype information and additional individuals genotyped for the HapA variants. The additional genotypes explain why there are 762 cases used in the analysis of Dr. Manolescu compared to the 713 included in the analysis of

Dr. Helgadottir. Updated phenotype information on the history of cardiovascular diseases for individuals used as controls in the analysis of Dr. Helgadottir led to the exclusion of some individuals from the control group, as these individuals reported either a history of coronary artery disease or related diseases.

18. For the Durham cohort, additional genotyping of the HapA variants led to the inclusion of additional 484 cases and 250 controls in the analysis of Dr. Manolescu, compared to the analysis of Dr. Helgadottir.

19. Thus, the difference between the analysis presented in the two declarations only arises as each one is prepared using the most updated and accurate data available to deCODE at the time when each analysis was done. Moreover, while these differences in datasets did result in different numerical calculations, the conclusion about the genetic correlation to be drawn in each instance was the same and statistically significant.

V. CERTIFICATION

20. I further declare that all statements made herein of my own knowledge are true, that all statements made on information and belief are believed to be true, and that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both (18 U.S.C. § 1001), and may jeopardize the validity of the application or any patent issuing thereon.

Dated: 15/5/2008

Gudmar Thorleifsson

Principal Investigator/Program Director *(Last, first, middle)*:

BIOGRAPHICAL SKETCH

NAME	POSITION TITLE
Thorleifsson, Gudmar	**Research Scientist**

EDUCATION/TRAINING

INSTITUTION AND LOCATION	DEGREE	YEAR(s)	FIELD OF STUDY
University of Iceland, Reykjavik, Iceland	B.Sc.	1984-87	Theoretical physics
University of Copenhagen, Copenhagen, Denmark	Cand. Scient	1987-92	High-energy physics
University of Copenhagen, Copenhagen, Denmark	Ph.D.	1992-94	High-energy physics
University of Syracuse, Syracuse, NY, USA		1994-96	Post-doc in physics

A. Positions

1992-94	Research assistant, University of Copenhagen, Copenhagen, Denmark
1994-96	Research Associate, Syracuse University, NY, USA
1996-99	Research Associate, University of Bielefeld, Bielefeld, Germany
1999-present	Research Scientist, deCode Genetics, Reykjavik, Iceland

B. Fellowships and Scholarships

1989-90	Danish Government Scholarship
1990	Carlsberg Foundation Scholarship
1992	University of Copenhagen Introduktionstipendium
1992-94	Carlsberg Foundation Candidatstipendium
1997-99	Alexander von Humboldt Research Fellowship
1999-x	Marie Curie (EC) Research Grant

C. Publications

1) G.Thorleifsson and P.H.Damgaard, The Kadanoff Lower-Bound Variational Renormalization Group applied to an SU(2) Lattice Spin Model, J. of Phys. **A23** (1990) 5863-5877.
2) G.Thorleifsson and P.H.Damgaard, The XY Model and Kadanoff's Variational Real-Space Renormalization Group, Physica **A178** (1991) 149-167.
3) P.H.Damgaard and G.Thorleifsson, Chaotic Renormalization Group Trajectories, Phys. Rev. **A44** (1991) 2738-2741.
4) J.Christensen, G.Thorleifsson, P.H.Damgaard and J.Wheater, Thermodynamics of SU(3) Lattice Gauge Theory in (2+1)-Dimension, Nucl. Phys. **B374** (1992) 225-248.
5) J.Christensen, G.Thorleifsson, P.H.Damgaard and J.Wheater, Three-Dimensional Deconfinement Transitions and Conformal Symmetry, Phys. Lett. **B276** (1992) 472-478.
6) J.Ambjorn, B.Durhuus, T.Jonsson and G.Thorleifsson, Matter Fields with c>1 coupled to 2D-Gravity, Nucl. Phys. **B398** (1993) 568-592.
7) J.Ambjorn, A.Sedrakyan and G.Thorleifsson, The 3D Ising Model represented as Random Surfaces, Phys. Lett. **B303** (1993) 327-333.
8) J.Ambjorn, S.Jain and G.Thorleifsson, Baby Universes in 2D Quantum Gravity, Phys. Lett. **B307** (1993) 34-39.
9) J.Ambjorn and G.Thorleifsson, A Universal Fractal Structure of 2D Quantum Gravity for c>1, Phys. Lett. **B323** (1994) 7-12.
10) J.Ambjorn, K.Farakos, S.Hands, G.Koutsoumbas and G.Thorleifsson, Level Crossing for Hot Sphalerons, Nucl. Phys. **B425** (1994) 39-66.
11) J.Ambjorn and G.Thorleifsson, Scaling with a modified Wilson Action which suppresses Z(2) Artifacts in SU(2) Lattice Gauge Theories, Phys. Rev. **D50** (1994) 4715-4717.
12) J.Ambjorn, G.Thorleifsson and M.Wexler, New Critical Phenomena in 2D Quantum Gravity, Nucl. Phys. **B439** (1995) 187-204.
13) S.M.Catterall, G.Thorleifsson, M.J.Bowick and V.John, Scaling and the Fractal Geometry of Two-Dimensional Quantum Gravity, Phys. Lett. **B354** (1995) 58-68.

14) S.M.Catterall, J.Kogut, R.Renken and G.Thorleifsson, Baby Universes in 4D Dynamical Triangulations, Phys. Lett. **B366** (1996) 72-76.
15) G.Thorleifsson and S.M.Catterall, A Real-Space Renormalization Group for Random Surfaces, Nucl. Phys. **B461** (1996) 350-370.
16) S.M.Catterall, G.Thorleifsson, J.Kogut and R.Renken, Singular Vertices and the Triangulation Space of the D-Sphere, Nucl. Phys. **B468** (1996) 263-276.
17) M.Bowick, S.M.Catterall, M.Falcioni, G.Thorleifsson and K.Anagnostopoulos, The Flat Phase of Crystalline Membranes, J. de Phys. **I** (France) **6** (1996) 1321-1345.
18) M.J.Bowick, S.M.Catterall, and G.Thorleifsson, Minimal Dynamical Triangulations of Random Surfaces, Phys. Lett. **B391** (1997) 305-309.
19) J.Ambjorn, K.N.Anagnostopoulos, U.Magnea and G.Thorleifsson, Geometrical Interpretation of the KPZ Exponents, Phys. Lett. **B388** (1996) 713-719.
20) M.J.Bowick, V.John and G.Thorleifsson, The Hausdorff Dimension of Surfaces in Two—Dimensional Quantum Gravity coupled to Unitary Minimal Matter, Phys. Lett. **B403** (1997) 197-202.
21) M.Falcioni, M.J.Bowick, E.Guitter and G.Thorleifsson, The Poisson Ratio of Crystalline Surfaces, Europhysics Letters **38**(1) (1997) 67-72.
22) M.J.Bowick, M.Falcioni and G.Thorleifsson, Numerical observation of a Tubular Phase in Anisotropic Membranes, Phys. Rev. Lett. **79** (1997) 885-889.
23) G.Thorleifsson and M.Falcioni, Improved Algorithm for Simulating Crystalline Membranes, Comp. Phys. Comm. **109** (1998) 161.
24) S.Bilke, Z.Burda, A.Krzywicki, B.Petersson, J.Tabaczek and G.Thorleifsson, 4d Simplicial Quantum Gravity Interacting with Gauge Matter Fields, Phys. Lett. **B418** (1998) 266-272.
25) K.N.Anagnostopoulos, P.Bialas and G.Thorleifsson, The Ising Model on a Quenched Ensemble of c=-5 Gravity Graphs, J. of Stat. Phys. **94** 3/4 (1999) 321-345.
26) S.Bilke, Z.Burda, A.Krzywicki, B.Petersson, J.Tabaczek and G.Thorleifsson, 4d Simplicial Quantum Gravity: Matter Fields and the Corresponding Effective Action, Phys. Lett. **B432** (1998) 279-286.
27) G.Thorleifsson, Three-Dimensional Simplicial Gravity and Degenerate Triangulations, Nucl. Phys. **B538** (1999) 278-294.
28) E.B.Gregory, S.M.Catterall and G.Thorleifsson, Monte Carlo Renormalization of 2D Simplical Quantum Gravity Coupled to Gaussian Matter, Nucl. Phys. **B541** (1999) 289-304.
29) S.Bilke and G.Thorleifsson, Simulating Four-Dimensional Simplicial Gravity using Degenerate Triangulations, Phys. Rev. **D59** (1999) 124008.
30) G.Thorleifsson, P.Bialas and B.Petersson, The Weak-Coupling Limit of Simplical Quantum Gravity, Nucl. Phys. **B550** (1999) 465-491.
31) M.Bowick, A.Cacciuto, G.Thorleifsson and A.Travesset, Universal Negative Poisson Ratio of Self-Avoiding Fixed-Connectivity Membranes, Phys. Rev. Lett. **87** (2001) 148103.
32) M.Bowick, A.Cacciuto, G.Thorleifsson and A.Travesset, Universality classes of self-avoiding fixed-connectivity membranes, European Physical Journal **E5** (2): 149-160 MAY 2001
33) Grant SF, Thorleifsson G, Frigge ML, Thorsteinsson J, Gunnlaugsdottir B, Geirsson AJ, Gudmundsson M, Vikingsson A, Erlendsson K, Valsson J, Jonsson H, Gudbjartsson DF, Stefansson K, Gulcher JR, Steinsson K., The Inheritance of Rheumatoid Arthritis in Iceland, Arthritis & Rheumatism, **44** (2001) 2247-2254.
34) Kristjansson K, Manolescu A, Kristinsson A, Hardarson T, Knudsen H, Ingason S, Thorleifsson G, Frigge ML, Kong A, Gulcher JR, Stefansson K., Linkage of Essential Hypertension to Chromosome 18q, Hypertension, **39** (2002) 1044-1049.
35) Reynisdottir I, Thorleifsson G, Benediktsson R, Sigurdsson G, Emilsson V, Einarsdottir AS, Hjorleifsdottir EE, Orlygsdottir GT, Bjornsdottir GT, Saemundsdottir J, Halldorsson S, Hrafnkelsdottir S, Sigurjonsdottir SB, Steinsdottir S, Martin M, Kochan JP, Rhees BK, Grant SF, Frigge ML, Kong A, Gudnason V, Stefansson K, Gulcher JR., Localization of a Susceptibility Gene for Type 2 Diabetes to Chromosome 5q32-q35.2, Am. J. Hum. Genet. **73** (2003) 323-35.
36) S.Gretarsdottir, G.Thorleifsson, S.T.Reynisdottir, A.Manolescu, S.Jonsdottir, T.Jonsdottir, T.Gudmundsdottir, S.M.Bjarnadottir, O.B.Einarsson, H.M.Gudjonsdottir, M.Hawkins, G.Gudmundsson, H.Gudmundsdottir, H.Andrason, A.S.Gudmundsdottir, M.Sigurdardottir, T.T.Chou, J.Nahmias, S.Goss, S.Sveinbjörnsdottir, E.M.Valdimarsson, F.Jakobsson, U.Agnarsson, V.Gudnason, G.Thorgeirsson, J.Fingerle, M.Gurney, D.Gudbjartsson, M.L.Frigge, A.Kong, K.Stefansson and J.R.Gulcher, The Gene Encoding Phosphodiesterase 4D Confers Risk of Ischemic Stroke, Nature Genetics, Vol **35**(2) (2003) 131-138.
37) Helgadottir A, Manolescu A, Thorleifsson G, Gretarsdottir S, Jonsdottir H, Thorsteinsdottir U, Samani NJ, Gudmundsson G, Grant SFA, Thorgeirsson G, Sveinbjornsdottir S, Valdimarsson EM, Matthiasson SE, Johannsson H, Gudmundsdottir O, Gurney ME, Sainz J, Thorhallsdottir M, Andresdottir M, Frigge ML, Topol EJ, Kong A, Gudnason V, Hakonarson H, Gulcher JR, Stefansson K, The gene encoding 5-lipoxygenase activating protein confers risk of myocardial infarction and stroke, Nature Genetics Vol **36**(3) (1004) 233-239
38) Kong A, Barnard J, Gudbjartsson DF, Thorleifsson G, Jonsdottir G, Sigurdardottir S, Richardsson B, Jonsdottir J, Thorgeirsson T, Frigge ML, Lamb NE, Sherman S, Gulcher JR, Stefansson K, Recombination rate and reproductive success in humans, Nature Genetics **36**(11) (2004) 1203-1206.

39) Helgadottir A, Gretarsdottir S, St.Clair D, Manolescu A, Cheung J, Thorleifsson G, Pasdar A, Granf SFA, Whalley, LJ, Hakonarson H, Thorsteinsdottir U, Kong A, Gulcher J, Stefansson K, MacLeod J, Association between the gene encoding 5-lipoxygenasa-activating protein and stroke replicated in a Scottish population, Am. J. Hum. Genet. **76** (2005) 505-509.

40) H Stefansson, A Helgason, G Thorleifsson, V Steinthorsdottir, G Masson, J Barnard, A Baker, A Jonasdottir, A Ingason, VG Gudnadottir, N Desnica, A Hicks, A Gylfason, DF Gudbjartsson, GM Jonsdottir, J Sainz, K Agnarsson, B Birgisdottir, S Ghosh, A Olafsdottir, JB Cazier, K Kristjansson, ML Frigge, TE Thorgeirsson, JR Gulcher, A Kong and K Stefansson, A common inversion under selection in Europeans, Nature Genetics Vol **37**(3) (2005) 129-137.

41) Gretarsdottir S, Gulcher J, Thorleifsson G, Kong A, Stefansson K. Comment on the phosphodiesterase 4D replication study by Bevan et al. Stroke. 2005 Sep 36(9):1824.

42) 22: Related Articles, LinksHelgadottir A, Manolescu A, Helgason A, Thorleifsson G, Thorsteinsdottir U, Gudbjartsson DF, Gretarsdottir S, Magnusson KP, Gudmundsson G, Hicks A, Jonsson T, Grant SF, Sainz J, O'Brien SJ, Sveinbjornsdottir S, Valdimarsson EM, Matthiasson SE, Levey AI, Abramson JL, Reilly MP, Vaccarino V, Wolfe ML, Gudnason V, Quyyumi AA, Topol EJ, Rader DJ, Thorgeirsson G, Gulcher JR, Hakonarson H, Kong A, Stefansson K. A variant of the gene encoding leukotriene A4 hydrolase confers ethnicity-specific risk of myocardial infarction. Nat Genet. 2006 Jan;38(1):68-74.

43) Magnusson KP, Duan S, Sigurdsson H, Petursson H, Yang Z, Zhao Y, Bernstein PS, Ge J, Jonasson F, Stefansson E, Helgadottir G, Zabriskie NA, Jonsson T, Björnsson A, Thorlacius T, Jonsson PV, Thorleifsson G, Kong A, Stefansson H, Zhang K, Stefansson K, Gulcher JR. CFH Y402H confers similar risk of soft drusen and both forms of advanced AMD. PLoS Med. 2006 Jan;3(1):e5.

44) Sainz J, Rovensky P, Gudjonsson SA, Thorleifsson G, Stefansson K, Gulcher JR. Segmental duplication density decrease with distance to human-mouse breaks of synteny. Eur J Hum Genet. 2006 Feb;14(2):216-21.

45) Grant SF, Thorleifsson G, Reynisdottir I, Benediktsson R, Manolescu A, Sainz J, Helgason A, Stefansson H, Emilsson V, Helgadottir A, Styrkarsdottir U, Magnusson KP, Walters GB, Palsdottir E, Jonsdottir T, Gudmundsdottir T, Gylfason A, Saemundsdottir J, Wilensky RL, Reilly MP, Rader DJ, Bagger Y, Christiansen C, Gudnason V, Sigurdsson G, Thorsteinsdottir U, Gulcher JR, Kong A, Stefansson K. Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. Nat Genet. 2006 Mar;38(3):320-3.

46) Amundadottir LT, Sulem P, Gudmundsson J, Helgason A, Baker A, Agnarsson BA, Sigurdsson A, Benediktsdottir KR, Cazier JB, Sainz J, Jakobsdottir M, Kostic J, Magnusdottir DN, Ghosh S, Agnarsson K, Birgisdottir B, Le Roux L, Olafsdottir A, Blondal T, Andresdottir M, Gretarsdottir OS, Bergthorsson JT, Gudbjartsson D, Gylfason A, Thorleifsson G, Manolescu A, Kristjansson K, Geirsson G, Isaksson H, Douglas J, Johansson JE, Bälter K, Wiklund F, Montie JE, Yu X, Suarez BK, Ober C, Cooney KA, Gronberg H, Catalona WJ, Einarsson GV, Barkardottir RB, Gulcher JR, Kong A, Thorsteinsdottir U, Stefansson K. A common variant associated with prostate cancer in European and African populations. Nat Genet. 2006 Jun;38(6):652-8.

47) Helgason A, Pálsson S, Thorleifsson G, Grant SF, Emilsson V, Gunnarsdottir S, Adeyemo A, Chen Y, Chen G, Reynisdottir I, Benediktsson R, Hinney A, Hansen T, Andersen G, Borch-Johnsen K, Jorgensen T, Schäfer H, Faruque M, Doumatey A, Zhou J, Wilensky RL, Reilly MP, Rader DJ, Bagger Y, Christiansen C, Sigurdsson G, Hebebrand J, Pedersen O, Thorsteinsdottir U, Gulcher JR, Kong A, Rotimi C, Stefánsson K. Refining the impact of TCF7L2 gene variants on type 2 diabetes and adaptive evolution. Nat Genet. 2007 Feb;39(2):218-25.

48) Steinthorsdottir V, Thorleifsson G, Reynisdottir I, Benediktsson R, Jonsdottir T, Walters GB, Styrkarsdottir U, Gretarsdottir S, Emilsson V, Ghosh S, Baker A, Snorradottir S, Bjarnason H, Ng MC, Hansen T, Bagger Y, Wilensky RL, Reilly MP, Adeyemo A, Chen Y, Zhou J, Gudnason V, Chen G, Huang H, Lashley K, Doumatey A, So WY, Ma RC, Andersen G, Borch-Johnsen K, Jorgensen T, van Vliet-Ostaptchouk JV, Hofker MH, Wijmenga C, Christiansen C, Rader DJ, Rotimi C, Gurney M, Chan JC, Pedersen O, Sigurdsson G, Gulcher JR, Thorsteinsdottir U, Kong A, Stefansson K. A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. Nat Genet. 2007 Jun;39(6):770-5.

49) Lyon HN, Emilsson V, Hinney A, Heid IM, Lasky-Su J, Zhu X, Thorleifsson G, Gunnarsdottir S, Walters GB, Thorsteinsdottir U, Kong A, Gulcher J, Nguyen TT, Scheraga A, Pfeufer A, Meitinger T, Brönner G, Rief W, Soto-Quiros ME, Avila L, Klanderman B, Raby BA, Silverman EK, Weiss ST, Laird N, Ding X, Groop L, Tuomi T, Isomaa B, Bengtsson K, Butler JL, Cooper RS, Fox CS, O'Donnell CJ, Vollmert C, Celedón JC, Wichmann HE, Hebebrand J, Stefansson K, Lange C, Hirschhorn JN. The association of a SNP upstream of INSIG2 with body mass index is reproduced in several but not all cohorts. PLoS Genet. 2007 Apr 27;3(4):e61.

50) Helgadottir A, Thorleifsson G, Manolescu A, Gretarsdottir S, Blondal T, Jonasdottir A, Jonasdottir A, Sigurdsson A, Baker A, Palsson A, Masson G, Gudbjartsson DF, Magnusson KP, Andersen K, Levey AI, Backman VM, Matthiasdottir S, Jonsdottir T, Palsson S, Einarsdottir H, Gunnarsdottir S, Gylfason A, Vaccarino V, Hooper WC, Reilly MP, Granger CB, Austin H, Rader DJ, Shah SH, Quyyumi AA, Gulcher JR, Thorgeirsson G, Thorsteinsdottir U, Kong A, Stefansson K. A common variant on chromosome 9p21 affects the risk of myocardial infarction. Science. 2007 Jun 8;316(5830):1491-3.

51) Gudbjartsson DF, Arnar DO, Helgadottir A, Gretarsdottir S, Holm H, Sigurdsson A, Jonasdottir A, Baker A, Thorleifsson G, Kristjansson K, Palsson A, Blondal T, Sulem P, Backman VM, Hardarson GA, Palsdottir E, Helgason A, Sigurjonsdottir R, Sverrisson JT, Kostulas K, Ng MC, Baum L, So WY, Wong KS, Chan JC, Furie KL, Greenberg SM, Sale M, Kelly P, MacRae CA, Smith EE, Rosand J, Hillert J, Ma RC, Ellinor PT, Thorgeirsson G, Gulcher JR, Kong A, Thorsteinsdottir U, Stefansson K. Variants conferring risk of atrial fibrillation on chromosome 4q25. Nature. 2007 Jul 19;448(7151):353-7.

52) Gudmundsson J, Sulem P, Steinthorsdottir V, Bergthorsson JT, Thorleifsson G, Manolescu A, Rafnar T, Gudbjartsson D, Agnarsson BA, Baker A, Sigurdsson A, Benediktsdottir KR, Jakobsdottir M, Blondal T, Stacey SN, Helgason A, Gunnarsdottir S, Olafsdottir A, Kristinsson KT, Birgisdottir B, Ghosh S, Thorlacius S, Magnusdottir D, Stefansdottir G, Kristjansson K, Bagger Y, Wilensky RL, Reilly MP, Morris AD, Kimber CH, Adeyemo A, Chen Y, Zhou J, So WY, Tong PC, Ng MC, Hansen T, Andersen G, Borch-Johnsen K, Jorgensen T, Tres A, Fuertes F, Ruiz-Echarri M, Asin L, Saez B, van Boven E, Klaver S, Swinkels DW, Aben KK, Graif T, Cashy J, Suarez BK, van Vierssen Trip O, Frigge ML, Ober C, Hofker MH, Wijmenga C, Christiansen C, Rader DJ, Palmer CN, Rotimi C, Chan JC, Pedersen O, Sigurdsson G, Benediktsson R, Jonsson E, Einarsson GV, Mayordomo JI, Catalona WJ, Kiemeney LA, Barkardottir RB, Gulcher JR, Thorsteinsdottir U, Kong A, Stefansson K. Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes. Nat Genet. 2007 Aug;39(8):977-83.

53) Kostulas K, Gretarsdottir S, Kostulas V, Manolescu A, Helgadottir A, Thorleifsson G, Gudmundsson LJ, Thorsteinsdottir U, Gulcher JR, Stefansson K, Hillert J. PDE4D and ALOX5AP genetic variants and risk for Ischemic Cerebrovascular Disease in Sweden. J Neurol Sci. 2007 Dec 15;263(1-2):113-7.

54) Thorleifsson G, Magnusson KP, Sulem P, Walters GB, Gudbjartsson DF, Stefansson H, Jonsson T, Jonasdottir A, Jonasdottir A, Stefansdottir G, Masson G, Hardarson GA, Petursson H, Arnarsson A, Motallebipour M, Wallerman O, Wadelius C, Gulcher JR, Thorsteinsdottir U, Kong A, Jonasson F, Stefansson K. Common sequence variants in the LOXL1 gene confer susceptibility to exfoliation glaucoma. Science. 2007 Sep 7;317(5843):1397-400.

55) Sulem P, Gudbjartsson DF, Stacey SN, Helgason A, Rafnar T, Magnusson KP, Manolescu A, Karason A, Palsson A, Thorleifsson G, Jakobsdottir M, Steinberg S, Pálsson S, Jonasson F, Sigurgeirsson B, Thorisdottir K, Ragnarsson R, Benediktsdottir KR, Aben KK, Kiemeney LA, Olafsson JH, Gulcher J, Kong A, Thorsteinsdottir U, Stefansson K. Genetic determinants of hair, eye and skin pigmentation in Europeans. Nat Genet. 2007 Dec;39(12):1443-52.

56) Helgadottir A, Thorleifsson G, Magnusson KP, Grétarsdottir S, Steinthorsdottir V, Manolescu A, Jones GT, Rinkel GJ, Blankensteijn JD, Ronkainen A, Jääskeläinen JE, Kyo Y, Lenk GM, Sakalihasan N, Kostulas K, Gottsäter A, Flex A, Stefansson H, Hansen T, Andersen G, Weinsheimer S, Borch-Johnsen K, Jorgensen T, Shah SH, Quyyumi AA, Granger CB, Reilly MP, Austin H, Levey AI, Vaccarino V, Palsdottir E, Walters GB, Jonsdottir T, Snorradottir S, Magnusdottir D, Gudmundsson G, Ferrell RE, Sveinbjornsdottir S, Hernesniemi J, Niemelä M, Limet R, Andersen K, Sigurdsson G, Benediktsson R, Verhoeven EL, Teijink JA, Grobbee DE, Rader DJ, Collier DA, Pedersen O, Pola R, Hillert J, Lindblad B, Valdimarsson EM, Magnadottir HB, Wijmenga C, Tromp G, Baas AF, Ruigrok YM, van Rij AM, Kuivaniemi H, Powell JT, Matthiasson SE, Gulcher JR, Thorgeirsson G, Kong A, Thorsteinsdottir U, Stefansson K. The same sequence variant on 9p21 associates with myocardial infarction, abdominal aortic aneurysm and intracranial aneurysm. Nat Genet. 2008 Feb;40(2):217-24.

57) Kong A, Thorleifsson G, Stefansson H, Masson G, Helgason A, Gudbjartsson DF, Jonsdottir GM, Gudjonsson SA, Sverrisson S, Thorlacius T, Jonasdottir A, Hardarson GA, Palsson ST, Frigge ML, Gulcher JR, Thorsteinsdottir U, Stefansson K. Sequence variants in the RNF212 gene associate with genome-wide recombination rate. Science. 2008 Mar 7;319(5868):1398-401.

58) Gudmundsson J, Sulem P, Rafnar T, Bergthorsson JT, Manolescu A, Gudbjartsson D, Agnarsson BA, Sigurdsson A, Benediktsdottir KR, Blondal T, Jakobsdottir M, Stacey SN, Kostic J, Kristinsson KT, Birgisdottir B, Ghosh S, Magnusdottir DN, Thorlacius S, Thorleifsson G, Zheng SL, Sun J, Chang BL, Elmore JB, Breyer JP, McReynolds KM, Bradley KM, Yaspan BL, Wiklund F, Stattin P, Lindström S, Adami HO, McDonnell SK, Schaid DJ, Cunningham JM, Wang L, Cerhan JR, St Sauver JL, Isaacs SD, Wiley KE, Partin AW, Walsh PC, Polo S, Ruiz-Echarri M, Navarrete S, Fuertes F, Saez B, Godino J, Weijerman PC, Swinkels DW, Aben KK, Witjes JA, Suarez BK, Helfand BT, Frigge ML, Kristjansson K, Ober C, Jonsson E, Einarsson GV, Xu J, Gronberg H, Smith JR, Thibodeau SN, Isaacs WB, Catalona WJ, Mayordomo JI, Kiemeney LA, Barkardottir RB, Gulcher JR, Thorsteinsdottir U, Kong A, Stefansson K. Common sequence variants on 2p15 and Xp11.22 confer susceptibility to prostate cancer. Nat Genet. 2008 Mar;40(3):281-3.

59) Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, Carlson S, Helgason A, Walters GB, Gunnarsdottir S, Mouy M, Steinthorsdottir V, Eiriksdottir GH, Bjornsdottir G, Reynisdottir I, Gudbjartsson D, Helgadottir A, Jonasdottir A, Jonasdottir A, Styrkarsdottir U, Gretarsdottir S, Magnusson KP, Stefansson H, Fossdal R, Kristjansson K, Gislason HG, Stefansson T, Leifsson BG, Thorsteinsdottir U, Lamb JR, Gulcher JR, Reitman ML, Kong A, Schadt EE, Stefansson K. Genetics of gene expression and its effect on disease. Nature. 2008 Mar 27;452(7186):423-8.

60) Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, Hu T, de Bakker PI, Abecasis GR, Almgren P, Andersen G, Ardlie K, Boström KB, Bergman RN, Bonnycastle LL, Borch-Johnsen K, Burtt NP, Chen H, Chines PS, Daly MJ, Deodhar P, Ding CJ, Doney AS, Duren WL, Elliott KS, Erdos MR, Frayling TM, Freathy RM, Gianniny L, Grallert H, Grarup N, Groves CJ, Guiducci C, Hansen T, Herder C, Hitman GA, Hughes TE, Isomaa B, Jackson AU, Jørgensen T, Kong A, Kubalanza K, Kuruvilla FG, Kuusisto J, Langenberg C, Lango H, Lauritzen T, Li Y, Lindgren CM, Lyssenko V, Marvelle AF, Meisinger C, Midthjell K, Mohlke KL, Morken MA, Morris AD, Narisu N, Nilsson P, Owen KR, Palmer CN, Payne F, Perry JR, Pettersen E, Platou C, Prokopenko I, Qi L, Qin L, Rayner NW, Rees M, Roix JJ, Sandbaek A, Shields B, Sjögren M, Steinthorsdottir V, Stringham HM, Swift AJ, Thorleifsson G, Thorsteinsdottir U, Timpson NJ, Tuomi T, Tuomilehto J, Walker M, Watanabe RM, Weedon MN, Willer CJ; Wellcome Trust Case Control Consortium, Illig T, Hveem K, Hu FB, Laakso M, Stefansson K, Pedersen O, Wareham NJ, Barroso I, Hattersley AT, Collins FS, Groop L, McCarthy MI, Boehnke M, Altshuler D. Meta-analysis

of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat Genet. 2008 May;40(5):638-45.

61) Thorgeirsson TE, Geller F, Sulem P, Rafnar T, Wiste A, Magnusson KP, Manolescu A, Thorleifsson G, Stefansson H, Ingason A, Stacey SN, Bergthorsson JT, Thorlacius S, Gudmundsson J, Jonsson T, Jakobsdottir M, Saemundsdottir J, Olafsdottir O, Gudmundsson LJ, Bjornsdottir G, Kristjansson K, Skuladottir H, Isaksson HJ, Gudbjartsson T, Jones GT, Mueller T, Gottsäter A, Flex A, Aben KK, de Vegt F, Mulders PF, Isla D, Vidal MJ, Asin L, Saez B, Murillo L, Blondal T, Kolbeinsson H, Stefansson JG, Hansdottir I, Runarsdottir V, Pola R, Lindblad B, van Rij AM, Dieplinger B, Haltmayer M, Mayordomo JI, Kiemeney LA, Matthiasson SE, Oskarsson H, Tyrfingsson T, Gudbjartsson DF, Gulcher JR, Jonsson S, Thorsteinsdottir U, Kong A, Stefansson K. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. Nature. 2008 Apr 3;452(7187):638-42.

62) Lasky-Su J, Lyon HN, Emilsson V, Heid IM, Molony C, Raby BA, Lazarus R, Klanderman B, Soto-Quiros ME, Avila L, Silverman EK, Thorleifsson G, Thorsteinsdottir U, Kronenberg F, Vollmert C, Illig T, Fox CS, Levy D, Laird N, Ding X, McQueen MB, Butler J, Ardlie K, Papoutsakis C, Dedoussis G, O'Donnell CJ, Wichmann HE, Celedón JC, Schadt E, Hirschhorn J, Weiss ST, Stefansson K, Lange C. On the replication of genetic associations: timing can be everything! Am J Hum Genet. 2008 Apr;82(4):849-58.

63) Gudbjartsson DF, Walters GB, Thorleifsson G, Stefansson H, Halldorsson BV, Zusmanovich P, Sulem P, Thorlacius S, Gylfason A, Steinberg S, Helgadottir A, Ingason A, Steinthorsdottir V, Olafsdottir EJ, Olafsdottir GH, Jonsson T, Borch-Johnsen K, Hansen T, Andersen G, Jorgensen T, Pedersen O, Aben KK, Witjes JA, Swinkels DW, den Heijer M, Franke B, Verbeek AL, Becker DM, Yanek LR, Becker LC, Tryggvadottir L, Rafnar T, Gulcher J, Kiemeney LA, Kong A, Thorsteinsdottir U, Stefansson K. Many sequence variants affecting diversity of adult human height. Nat Genet. 2008 May;40(5):609-15.

Pearson TA & Manolio, TA, JAMA 299:1335-1344 (2008)

How to Interpret a Genome-wide Association Study

Thomas A. Pearson, MD, MPH, PhD
Teri A. Manolio, MD, PhD

Genome-wide association (GWA) studies use high-throughput genotyping technologies to assay hundreds of thousands of single-nucleotide polymorphisms (SNPs) and relate them to clinical conditions and measurable traits. Since 2005, nearly 100 loci for as many as 40 common diseases and traits have been identified and replicated in GWA studies, many in genes not previously suspected of having a role in the disease under study, and some in genomic regions containing no known genes. GWA studies are an important advance in discovering genetic variants influencing disease but also have important limitations, including their potential for false-positive and false-negative results and for biases related to selection of study participants and genotyping errors. Although these studies are clearly many steps removed from actual clinical use, and specific applications of GWA findings in prevention and treatment are actively being pursued, at present these studies mainly represent a valuable discovery tool for examining genomic function and clarifying pathophysiologic mechanisms. This article describes the design, interpretation, application, and limitations of GWA studies for clinicians and scientists for whom this evolving science may have great relevance.

JAMA. 2008;299(11):1335-1344 www.jama.com

In the past 2 years, there has been a dramatic increase in genomic discoveries involvng complex, non-Mendelian diseases, with nearly 100 loci for as many as 40 common diseases robustly identified and replicated in genome-wide association (GWA) studies (T.A.M.; unpublished data, 2008). These studies use high-throughput genotyping technologies to assay hundreds of thousands of the most common form of genetic variant, the single-nucleotide polymorphism (SNP), and relate these variants to diseases or health-related traits.[1] Nearly 12 million unique human SNPs have been assigned a reference SNP (rs) number in the National Center for Biotechnology Information's dbSNP database[2] and characterized as to specific alleles (alternate forms of the SNP), summary allele frequencies, and other genomic information.[3]

The GWA approach is revolutionary because it permits interrogation of the entire human genome at levels of resolution previously unattainable, in thousands of unrelated individuals, unconstrained by prior hypotheses regarding genetic associations with disease.[4] However, the GWA approach can also be problematic because the massive number of statistical tests performed presents an unprecedented potential for false-positive results, leading to new stringency in acceptable levels of statistical significance and requirements for replication of findings.[5]

The genome-wide, nonhypothesis-driven nature of GWA studies represents an important step beyond candidate gene studies, in which the high cost of genotyping had limited the number of variants assayed to several hundred at most. This required careful selection of variants to be studied, often based on imperfect understanding of the biologic pathways relating genes to disease.[6] Many such associations failed to be replicated in subsequent studies,[7,8] leading to calls for all genetic association reports to include documented replication of findings as a prerequisite for publication.[9,10]

For non-Mendelian conditions, GWA studies also represent a valuable advance over family-based linkage studies, in which multiply affected families are arduously assembled and inheritance patterns are related to several hundred markers throughout the genome. Family-based linkage studies, although successful in identifying genes of large effect in Mendelian diseases such as cystic fibrosis and neurofibromatosis, have had more limited success in common diseases like atherosclerosis and asthma.[11] Major limitations of linkage studies are relatively low power for complex disorders influenced by multiple genes, and the large size of the chromosomal regions shared among family members (often comprising hundreds of genes), in whom it can be difficult to narrow the

Author Affiliations: Office of Population Genomics, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland (Drs Pearson and Manolio); and Clinical and Translational Science Institute, University of Rochester Medical Center, Rochester, New York (Dr Pearson).
Corresponding Author: Teri A. Manolio, MD, PhD, Office of Population Genomics, National Human Genome Research Institute, 31 Center Dr, Room 4B-09, MSC 2154, Bethesda, MD 20892-2154 (manolio@nih.gov).

linkage signal sufficiently to identify a causative gene.

GWA studies build on the valuable lessons learned from candidate gene and family linkage studies, as well as the expanding knowledge of the relationships among SNP variants generated by the International HapMap Project,[12,13] to capture the great majority of common genetic differences among individuals and relate them to health and disease. These studies not only represent a powerful new tool for identification of genes influencing common diseases, but also use new terminologies (BOX 1), apply new models, and present new challenges in interpretation. GWA studies rely on the "common disease, common variant" hypothesis, which suggests that genetic influences on many common diseases will be at least partly attributable to a limited number of allelic variants present in more than 1% to 5% of the population.[14] Many important disease-causing variants may be rarer than this and are unlikely to be detected with this approach.

Although GWA discovery studies provide important clues to genomic function and pathophysiologic mechanisms, they are as yet many steps removed from actual clinical application. Nonetheless, they have gained considerable media attention and have the potential for generating queries from patients about whether to get tested for the "new gene for disease X" based on the latest report. In this article, we describe the design, interpretation, application, and limitations of GWA studies for clinicians and scientists for whom this evolving science may have great relevance.

Overview of GWA Studies

A GWA study is defined by the National Institutes of Health as a study of common genetic variation across the entire human genome designed to identify genetic associations with observable traits.[15] Although family linkage studies and studies comprising tens of thousands of gene-based SNPs also assay genetic variation across the genome,[16] the National Institutes of Health definition requires sufficient density and selection of genetic markers to capture a large proportion of the common variants in the study population, measured in enough individuals to provide sufficient power to detect variants of modest effect.

The present discussion focuses on studies attempting to assay at least 100 000 SNPs selected to serve as proxies for the largest possible number of SNPs.[12] The typical GWA study has 4 parts: (1) selection of a large number of individuals with the disease or trait of interest and a suitable comparison group; (2) DNA isolation, genotyping, and data review to ensure high genotyping quality; (3) statistical tests for associations between the SNPs passing quality thresholds and the disease/trait; and (4) replication of identified associations in an independent population sample or examination of functional implications experimentally.

Most of the roughly 100 GWA studies published by the end of 2007 were designed to identify SNPs associated with common diseases. However, the technique can also be used to identify genetic variants related to quantitative traits such as height[17] or electrocardiographic QT interval,[18] and to rank the relative importance of previously identified susceptibility genes, such as *APOE*e4* in Alzheimer disease[19] and *CARD15* and *IL23R* in Crohn disease.[20]

GWA studies can also demonstrate gene-gene interactions, or modification of the association of one genetic variant by another, as with *GAB2* and *APOE* in Alzheimer disease,[21] and can detect high-risk haplotypes or combinations of multiple SNPs within a single gene, as in exfoliation glaucoma[22] and atrial fibrillation.[23] These studies have also been used to identify SNPs associated with gene expression, either as confirmation of a phenotypic association, such as asthma and *ORMDL3* expression,[24] or more globally.[25] Thus, GWA studies have broader applications than those solely involving discovery of individual SNPs associated with discrete disease end points.

Study Designs Used in GWA

By far the most frequently used GWA study design to date has been the case-control design, in which allele frequencies in patients with the disease of interest are compared to those in a disease-free comparison group. These studies are often easier and less expensive to conduct than studies using other designs, especially if sufficient numbers of case and control participants can be assembled rapidly. This design also carries the most assumptions, which if not met, can lead to substantial biases and spurious associations (TABLE 1). The most important of these biases involve the selected, often unrepresentative nature of the study case participants, who are typically sampled from clinical sources and thus may not include fatal, mild, or silent cases not coming to clinical attention; and the lack of comparability of case and control participants, who may differ in important ways that could be related both to genetic risk factors and to disease outcomes.[26]

If well-established principles of epidemiologic design are followed, case-control studies can produce valid results that, especially for rare diseases, may not be obtainable in any other way. However, genetic association studies using case-control methodologies have often not always adhered to these principles. The often sharply abbreviated descriptions of case and control participants and lack of comparison of key characteristics in GWA reports[27] can make evaluation of potential biases and replication of findings quite difficult.[28]

The trio design includes the affected case participant and both of his or her parents.[29] Phenotypic assessment (classification of affected status) is performed only in the offspring and only affected offspring are included, but genotyping is performed in all 3 trio members. The frequency with which an allele is transmitted to an affected offspring from heterozygous parents is then estimated.[29] Under the null hy-

Box 1. Terms Frequently Used in Genome-wide Association Studies

Alleles
Alternate forms of a gene or chromosomal locus that differ in DNA sequence

Candidate gene
A gene believed to influence expression of complex phenotypes due to known biological and/or physiological properties of its products, or to its location near a region of association or linkage

Copy number variants
Stretches of genomic sequence of roughly 1 kb to 3 Mb in size that are deleted or are duplicated in varying numbers

False discovery rate[59,60]
Proportion of significant associations that are actually false positives

False-positive report probability[61]
Probability that the null hypothesis is true, given a statistically significant finding

Functional studies
Investigations of the role or mechanism of a genetic variant in causation of a disease or trait

Gene-environment interactions
Modification of gene-disease associations in the presence of environmental factors

Genome-wide association study
Any study of genetic variation across the entire human genome designed to identify genetic association with observable traits or the presence or absence of a disease, usually referring to studies with genetic marker density of 100 000 or more to represent a large proportion of variation in the human genome

Genotyping call rate
Proportion of samples or SNPs for which a specific allele SNP can be reliably identified by a genotyping method

Haplotype
A group of specific alleles at neighboring genes or markers that tend to be inherited together

HapMap[12,13]
Genome-wide database of patterns of common human genetic sequence variation among multiple ancestral population samples

Hardy Weinberg equilibrium
Population distribution of 2 alleles (with frequencies p and q) such that the distribution is stable from generation to generation and genotypes occur at frequencies of p^2, 2pq, and q^2 for the major allele homozygote, heterozygote, and minor allele homozygote, respectively

Linkage disequilibrium
Association between 2 alleles located near each other on a chromosome, such that they are inherited together more frequently than expected by chance

Mendelian disease
Condition caused almost entirely by a single major gene, such as cystic fibrosis or Huntington's disease, in which disease is manifested in only 1 (recessive) or 2 (dominant) of the 3 possible genotype groups

Minor allele
The allele of a biallelic polymorphism that is less frequent in the study population

Minor allele frequency
Proportion of the less common of 2 alleles in a population (with 2 alleles carried by each person at each autosomal locus) ranging from less than 1% to less than 50%

Modest effect
Association between a gene variant and disease or trait that is statistically significant but carries a small odds ratio (usually <1.5)

Non-Mendelian disease (also "common" or "complex" disease)
Condition influenced by multiple genes and environmental factors and not showing Mendelian inheritance patterns

Nonsynonymous SNP
A polymorphism that results in a change in the amino acid sequence of a protein (and therefore may affect the function of the protein)

Platform
Arrays or chips on which high-throughput genotyping is performed

Polymorphic
A gene or site with multiple allelic forms. The term *polymorphism* usually implies a minor allele frequency of at least 1%

Population attributable risk
Proportion of a disease or trait in the population that is due to a specific cause, such as a genetic variant

Population stratification (also "population structure")
A form of confounding in genetic association studies caused by genetic differences between cases and controls unrelated to disease but due to sampling them from populations of different ancestries

Power
A statistical term for the probability of identifying a difference between 2 groups in a study when a difference truly exists

Single-nucleotide polymorphism
Most common form of genetic variation in the genome, in which a single-base substitution has created 2 forms of a DNA sequence that differ by a single nucleotide

Tag SNP
A readily measured SNP that is in strong linkage disequilibrium with multiple other SNPs so that it can serve as a proxy for these SNPs on large-scale genotyping platforms

Trio
Genetic study design including an affected offspring and both parents

Abbreviation: SNP, single-nucleotide polymorphism.

pothesis of no association with disease, the transmission frequency for each allele of a given SNP will be 50%, but alleles associated with the disease will be transmitted in excess to the affected case individual. Because the trio design studies allele transmission from parents to offspring, it is not susceptible to population stratification, or genetic differences between case and control participants unrelated to disease but due to sampling them from populations of different ancestry.[30] A significant challenge of the trio design in GWA studies is its sensitivity to even small degrees of genotyping error,[4,31] which can distort transmission proportions between parents and offspring, especially for uncommon alleles. Therefore, standards for genotyping quality in trio studies may need to be more stringent than for other designs.

Cohort studies involve collecting extensive baseline information in a large number of individuals who are then observed to assess the incidence of disease in subgroups defined by

Table 1. Study Designs Used in Genome-wide Association Studies

	Case-Control	Cohort	Trio
Assumptions	Case and control participants are drawn from the same population Case participants are representative of all cases of the disease, or limitations on diagnostic specificity and representativeness are clearly specified Genomic and epidemiologic data are collected similarly in cases and controls Differences in allele frequencies relate to the outcome of interest rather than differences in background population between cases and controls	Participants under study are more representative of the population from which they are drawn Diseases and traits are ascertained similarly in individuals with and without the gene variant	Disease-related alleles are transmitted in excess of 50% to affected offspring from heterozygous parents
Advantages	Short time frame Large numbers of case and control participants can be assembled Optimal epidemiologic design for studying rare diseases	Cases are incident (developing during observation) and free of survival bias Direct measure of risk Fewer biases than case-control studies Continuum of health-related measures available in population samples not selected for presence of disease	Controls for population structure; immune to population stratification Allows checks for Mendelian inheritance patterns in genotyping quality control Logistically simpler for studies of children's conditions Does not require phenotyping of parents
Disadvantages	Prone to a number of biases including population stratification Cases are usually prevalent cases, may exclude fatal or short episodes, or mild or silent cases Overestimate relative risk for common diseases	Large sample size needed for genotyping if incidence is low Expensive and lengthy follow-up Existing consent may be insufficient for GWA genotyping or data sharing Requires variation in trait being studied Poorly suited for studying rare diseases	May be difficult to assemble both parents and offspring, especially in disorders with older ages of onset Highly sensitive to genotyping error

genetic variants. Although cohort studies are typically more expensive and take longer to conduct than case-control studies, they often include study participants who are more representative than clinical series of the population from which they are drawn, and they typically include a vast array of health-related characteristics and exposures for which genetic associations can be sought.[17,18] For these reasons, genome-wide genotyping has recently been added to cohort studies such as the Framingham Heart Study[32] and the Women's Health Study.[33]

Many GWA studies use multistage designs to reduce the number of false-positive results while minimizing the number of costly genome-wide scans performed and retaining statistical power.[4] Genome-wide scans are typically performed on an initial group of case and control participants and then a smaller number of associated SNPs is replicated in a second or third group of case and control participants (TABLE 2). Some studies begin with small numbers of participants in the initial scan but carry forward large numbers of SNPs to minimize false-negative results.[34] Other studies begin with more participants but carry forward a smaller proportion of associated SNPs.[35] Optimal proportions of study participants and SNPs in each phase have yet to be determined,[36] but carrying forward a small proportion (<5%) of stage 1 SNPs will often mean limiting the associations ultimately identified to those having a relatively large effect.[37]

Selection of Study Participants

Many genetic studies, whether GWA or otherwise, focus on case participants more likely to have a genetic basis for their disease, such as early-onset cases or those with multiple affected relatives. Misclassification of case participants can markedly reduce study power and bias study results toward no association, particularly when large numbers of unaffected individuals are misclassified as affected. For diseases that are difficult to diagnose reliably, ensuring that cases are truly affected (as by invasive testing or imaging), is probably more important than ensuring generalizability, although the limitations on diagnostic reliability and generalizability should be clearly described so that clinicians can judge the relevance to their patients.

The control participants should be drawn from the same population as the case participants and should be at risk to develop the disease and be detected in the study. Inclusion of women as controls in genetic association studies of diseases limited to men, for example, is problematic in that this approach adds individuals to the control group who had no chance of developing the disease (but might have done so had they also inherited a Y chromosome), thus mixing the controls with possible latent cases. This artificially reduces the differences in allele frequencies between cases and controls and limits the ability of the study to detect a true difference (ie, reduces study power).

If the disease is common, such as coronary heart disease or hypertension in the United States, efforts should be made to ensure that the controls are truly disease free. Some studies address this by using super-controls or persons at high risk but without even early evidence of disease, such as per-

sons with diabetes of long duration but without microalhuminuria in a study of diabetic nephropathy.[38] The success of recent GWA studies using control groups of questionable representativeness due to volunteer bias, such as the blood donor cohort in the Wellcome Trust Case-Control Consortium,[39] suggests that initial identification of SNPs associated with disease may be robust to these biases, especially given subsequent evidence of replication of these associations in studies using more traditional control groups.[40-42]

Of more concern may be the risk of false-negative findings, as many biases tend to reduce the magnitude of observed associations toward the null. Use of convenience controls such as blood donors, however, may also be problematic in examining potential modification of genetic associations by environmental exposures and sociocultural factors, and in the identification of less strongly associated SNPs.

A key component in articles reporting results in the epidemiology literature of observational study is an initial table comparing relevant characteristics of those with and without disease, allowing assessment of comparability and generalizability of the 2 groups. Such comparisons are infrequent in GWA studies,[38] but they are important because common diseases are typically influenced by multiple environmental (as well as genetic) factors. Important differences should be adjusted for in the analysis if possible, to avoid the risk of identifying genetic associations not with the disease of interest but with a confounding factor, such as smoking[43] or obesity.[44]

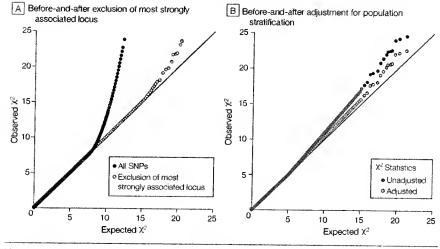
Confounding due to population stratification (also called population structure) has been cited as a major threat to the validity of genetic association studies, but its true importance is a matter of debate.[45,46] When variations occur in allele frequency between population subgroups, such as those defined by ethnicity or geographic origin, that in turn differ in their risk for disease, GWA studies may then falsely identify the subgroup-associated genes as relatcd to disease.[30] Population structure should be assessed and reported in GWA studies, typically by examining the distribution of test statistics generated from the thousands of association tests performed (eg, the χ^2 test) and assessing their deviation from the null distribution (that expected under the null hypothesis of no SNP associated with the trait) in a quantile-quantile or "Q-Q," plot (FIGURE 1). In these plots, observed association statistics or calculated P values for each SNP are ranked in order from smallest to largest and plotted against the values expected had they been sampled from a distribution of known form (such as the χ^2 distribution).[30] Deviations from the diagonal identity line suggest that either the assumed distribution is incorrect or that the sample contains values arising

Table 2. Examples of Multistage Designs in Genome-wide Association Studies[a]

	3-Stage Study[b]		4-Stage Study[c]	
Stage	Case Participants/ Control Participants	SNPs Analyzed	Case Participants/ Control Participants	SNPs Analyzed
1	400/400	500 000	2000/2000	100 000
2	4000/4000	25 000	2000/2000	1000
3	20 000/20 000	25	2000/2000	20
4			2000/2000	5

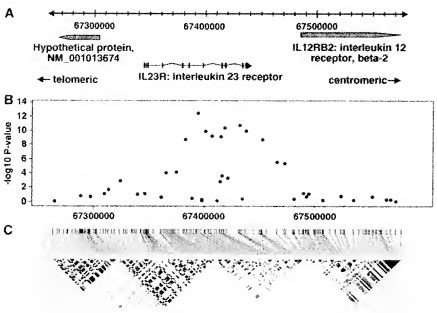
Abbreviation: SNP, single-nucleotide polymorphism.
[a] Based on hypothetical data.
[b] Five SNPs associated with disease.
[c] Two SNPs associated with disease.

Figure 1. Hypothetical Quantile-Quantile Plots in Genome-wide Association Studies



The Q-Q plot is used to assess the number and magnitude of observed associations between genotyped single-nucleotide polymorphisms (SNPs) and the disease or trait under study, compared to the association statistics expected under the null hypothesis of no association.[19] Observed association statistics (eg, χ^2 or t statistics) or $-\log_{10}$ P values calculated from them, are ranked in order from smallest to largest on the y-axis and plotted against the distribution that would be expected under the null hypothesis of no association on the x-axis. Deviations from the identity line suggest either that the assumed distribution is incorrect or that the sample contains values arising in some other manner, as by a true association.[19] A, Observed χ^2 statistics of all polymorphic SNPs (dark blue) in a hypothetical genome-wide association study of a complex disease vs. the expected null distribution (black line). The sharp deviation above an expected χ^2 value of approximately 8 could be due to a strong association of the disease with SNPs in a heavily genotyped region such as the major histocompatibility locus (MHC) on chromosome 6p21 in multiple sclerosis or rheumatoid arthritis.[19] Exclusion of SNPs from such a locus may leave a residual upward deviation (light blue) identifying more associated SNPs with higher observed χ^2 values (exceeding approximately 17) than expected under the null hypothesis. B, Observed (dark purple) vs expected (black line) χ^2 statistics for a hypothetical genome-wide association study of a complex disease. Deviation from the expected distribution is observed above an expected χ^2 of approximately 5. Inflation of observed statistics due to relatedness and potential population structure can be estimated by the method of genomic control.[19] Correction for this inflation by simple division reduces the unadjusted χ^2 statistics (dark purple) to the adjusted levels (light purple), showing deviation only above an expected χ^2 of approximately 15. The region between expected χ^2 of approximately 5 to approximately 15 is suggestive of broad differences in allele frequencies that are more likely due to population structure than disease susceptibility genes.

Figure 2. Associations in the *IL23R* Gene Region Identified by a Genome-wide Association Study of Inflammatory Bowel Disease



Genome-wide association studies frequently identify associations with many highly correlated single-nucleotide polymorphisms (SNPs) in a chromosomal region, due in part to linkage disequilibrium, among the SNPs. This can make it difficult to determine which SNP within a group is likely to be the causative or functional variant. A, Genomic locations of 2 genes, the interleukin 23 receptor (*IL23R*) and the interleukin 12 receptor, beta-2 (*IL12Rb2*), and a hypothetical protein, NM_001013674, between positions 62700000 and 67580000 of the short arm of chromosome 1 at region 1p31, are shown. B, The $-\log_{10} P$ values for association with inflammatory bowel disease are plotted for each SNP genotyped in the region; those reaching a prespecified value of $-\log_{10}$ of 7 or greater are presumed to show association with disease. Several strong associations, at $-\log_{10} P$ values or greater, are seen in the region just telomeric of position approximately 67400000 and extending just centromeric of position approximate 67450000. C, Pairwise linkage disequilibrium estimates between SNPs (measured as r^2) are plotted for the region. Higher r^2 values are indicated by darker shading. The region contains 4 "triangles" or "blocks" of linkage disequilibrium, 2 on either side of position 67400000 in the *IL23R* gene, another in the hypothetical protein telomeric of *IL23R*, and a fourth in the *IL12RB2* gene at the centromeric end of the region. The 2 *IL23R* linkage disequilibrium regions each contain SNPs associated with inflammatory bowel disease, while the *IL12RB2* region does not. Reproduced with permission from Duerr et al.[19]

in some other manner, as by a true association.[39]

Since the underlying assumption in GWA studies is that the vast majority of assayed SNPs are not associated with the trait, strong deviations from the null suggest either a very highly associated and heavily genotyped locus (Figure 1, A), or significant differences in population structure (Figure 1, B). Several effective statistical methods are available to correct for population structure and are a standard component of rigorous GWA analyses.[20,30]

Genotyping and Quality Control in GWA Studies

GWA studies rely on the typically strong associations among SNPs located near each other on a chromosome, which tend to be inherited together more often than expected by chance.[30] This nonrandom association is called linkage disequilibrium; alleles of SNPs in high linkage disequilibrium are almost always inherited together and can serve as proxies for each other. Their correlation with each other in the population is measured by the r^2 statistic, which is the proportion of variation of one SNP explained by the other, and ranges from 0 (no association) to 1 (perfect correlation).

Genomic coverage of GWA genotyping platforms (arrays or chips on which genotyping is performed) is often estimated by the percent of common SNPs having an r^2 of 0.8 or greater with at least 1 SNP on the platform.[13] Genotyping platforms comprising 500 000 to 1 000 000 SNPs have been estimated to capture 67% to 89% of common SNP variation in populations of European and Asian ancestry and 46% to 66% of variation in individuals of recent African ancestry.[13] Higher density platforms now also include probes for copy number variants that are not well tagged by SNPs. Copy number variants, in which stretches of genomic sequence are deleted or are duplicated in varying numbers, have gained increasing attention because of their apparent ubiquity and potential dosage effect on gene expression.[31] Newer genotyping platforms are increasingly being focused on capturing copy number variants, but other structural variants such as insertions, deletions, and inversions, remain difficult to assay.[32]

GWA studies frequently identify associations with multiple SNPs in a chromosomal region and display the association statistics by their genomic location on a portion of a chromosome (FIGURE 2). For ease of display, association statistics are typically shown as the $-\log_{10}$ of the P value (the probability of the observed association arising by chance alone), so that $P = .01$ would be plotted as "2" on the y-axis and $P = 10^{-7}$ as "7." Such displays also often plot a matrix of r^2 values for each pair of SNPs in the region, with larger r^2 values more intensely shaded. These plots can be used to identify linkage disequilibrium blocks containing SNPs associated with disease, allowing estimation of the independence of the SNP associations observed.[33]

Genotyping errors, especially if occurring differentially between cases and controls, are an important cause of spurious associations and must be diligently sought and corrected.[34] A number of quality control features should be applied both on a per-sample and a per-SNP basis. Checks on sample identity to avoid sample mix-ups should be described and a minimum rate of successfully genotyped SNPs per sample (usually 80%-90% of SNPs attempted) should be reported. Once samples failing these thresholds are removed, individual SNPs across the re-

Table 3. Association of Alleles and Genotypes of rs6983267 on Chromosome 8q24 With Colorectal Cancer[a]

	Number and Frequency of rs6983267 Alleles in Colorectal Cancer					Number and Frequency of rs6983267 Genotypes in Colorectal Cancer						
	C	T	χ^2 (1*df*)	*P* Value	OR	CC	CT	TT	χ^2 (2*df*)	*P* Value	OR	OR
Cases	875 (56.5)	675 (43.5)	24.8	6.3×10^{-7}	1.35[b]	250 (32.3)	375 (48.4)	150 (19.4)	24.5	4.7×10^{-6}	1.33[c]	1.81[d]
Controls	1860 (48.9)	1940 (51.1)				460 (24.2)	940 (49.4)	500 (26.3)				

Abbreviation: OR, odds ratio.
[a] Data are hypothetical, adapted from Tomlinson et al.[36]
[b] Denotes allelic odds ratio.
[c] Denotes heterozygote odds ratio.
[d] Denotes homozygote odds ratio.

maining samples are subjected to further checks or filters for probable genotyping errors, including: (1) the proportion of samples for which a SNP can be measured (the SNP call rate, typically >95%); (2) the minor allele frequency (often >1%, as rarer SNPs are difficult to measure reliably; (3) severe violations of Hardy-Weinberg equilibrium; (4) Mendelian inheritance errors in trio studies; and (5) concordance rates in duplicate samples (typically >99.5%).

Additional checks on genotyping quality should include careful visual inspection of genotype cluster plots, or intensity values generated by the genotyping assay to ensure that the strongest associations do not merely reflect genotyping artifact.[38,39] Genotyping the most strongly associated SNPs should also be confirmed using a different method.[38] Associations with any known "positive controls," such as *TCF7L2* in type 2 diabetes mellitus[33] or *HLA-DRB1* in rheumatoid arthritis,[47] should be reported to increase confidence in the consistency of findings with prior reports.

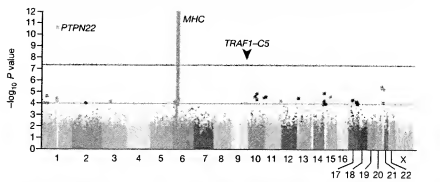
Analysis and Presentation of GWA Results

Associations with the 2 alleles of each SNP are tested in a relatively straightforward manner by comparing the frequency of each allele in cases and controls (TABLE 3). Because each individual carries 2 copies of each autosomal SNP, the frequency of each of 3 possible genotypes can also be compared (Table 3). Exploratory analyses may also include testing of different genetic models (dominant, recessive, or additive), although additive models, in which each copy of the allele is assumed to increase risk by the same amount, tend to be the most common (T.A.M.; unpublished data, 2008). Odds ratios of disease associated with the risk allele or genotype(s) can then be calculated and are typically modest, often in the range of 1.2 to 1.3. Many studies also calculate population attributable risk, classically defined as the proportion of disease in the population associated with a given risk factor (in this case, a genetic variant).[57]

Such estimates are nearly always inflated because odds ratios overestimate relative risks (especially for common diseases[58]) needed for population attributable risk calculations, and because odds ratios and allele frequencies in published reports have wide confidence intervals so that those selected by exceeding a specified threshold for statistical significance tend to be biased upwards, an effect of ascertainment known as the "winner's curse."[59] This exaggerated initial estimate of the odds ratio often leads to replication studies that lack sufficient sample size and power to replicate the association because larger samples are needed to detect smaller odds ratios.

Complexity in analysis emerges due to the multiple testing carried out in GWA studies, in that the association tests shown in Table 2 are repeated for each of the 100 000 to more than 1 million SNPs assayed (FIGURE 3). At the conventional $P < .05$ level of significance, an association study of 1 million SNPs will show 50 000 SNPs to be "associated" with disease, almost all

Figure 3. Genome-wide Association Findings in Rheumatoid Arthritis



Genome-wide association studies assume a priori hypotheses about candidate genes or regions that might be associated with disease; rather, they test single-nucleotide polymorphisms (SNPs) throughout the genome for possible evidence of genetic susceptibility. Associations plotted as $-\log_{10} P$ values for a genome-wide association study in 1522 cases with rheumatoid arthritis and 1850 controls, showing single data points for SNPs with $P < 10^{-4}$ (lower horizontal red line) for 22 autosomes and the X chromosome. The predefined level of significance, at 5×10^{-8} is shown with a horizontal blue line. SNPs at *PTPN22* on chromosome 1, the major histocompatibility complex (MHC) on chromosome 6, and the *TRAF1-C5* locus on chromosome 9 exceed this threshold. Reproduced with permission from Plenge et al.[47]

Box 2. Ten Basic Questions to Ask About a Genome-wide Association Study Report[a]

1. Are the cases defined clearly and reliably so that they can be compared with patients typically seen in clinical practice?
2. Are case and control participants demonstrated to be comparable to each other on important characteristics that might also be related to genetic variation and to the disease?
3. Was the study of sufficient size to detect modest odds ratios or relative risks (1.3-1.5)?
4. Was the genotyping platform of sufficient density to capture a large proportion of the variation in the population studied?
5. Were appropriate quality control measures applied to genotyping assays, including visual inspection of cluster plots and replication on an independent genotyping platform?
6. Did the study reliably detect associations with previously reported and replicated variants (known positives)?
7. Were stringent corrections applied for the many thousands of statistical tests performed in defining the *P* value for significant associations?
8. Were the results replicated in independent population samples?
9. Were the replication samples comparable in geographic origin and phenotype definition, and if not, did the differences extend the applicability of the findings?
10. Was evidence provided for a functional role for the gene polymorphism identified?

[a]For a more detailed description of interpretation of genome-wide association studies, see NCI/NHGRI Working Group on Replication in Association Studies.[28]

falsely positive and due to chance alone. The most common manner of dealing with this problem is to reduce the false-positive rate by applying the Bonferroni correction, in which the conventional *P* value is divided by the number of tests performed.[60] A 1 million SNP survey would thus use a threshold of $P < .05/10^6$, or 5×10^{-8}, to identify associations unlikely to have occurred by chance. This correction has been criticized as overly conservative because it assumes independent associations of each SNP with disease even though individual SNPs are known to be correlated to some degree due to linkage disequilibrium.

Other approaches have been proposed, including estimation of the false discovery rate or proportion of significant associations that are actually false positive associations,[61,62] false-positive report probability, or probability that the null hypothesis is true given a statistically significant finding,[63] and estimation of Bayes factors that incorporate the prior probability of association based on characteristics of the disease or the specific SNP.[39] To date, Bonferroni correction has generally been the most commonly used correction for multiple comparisons in GWA reports (T.A.M.; unpublished data, 2008).

Replication and Functional Studies

Given the major challenge of separating the many false-positive associations from the few true-positive associations with disease in GWA studies, an important strategy has been replication of results in independent samples.[28] This is typically included in a single GWA report as part of a multistage design[34,35] or may be reported separately.[39,64] Consensus criteria for replication have recently been published and include study of the same or very similar phenotype and population, and demonstration of a similar magnitude of effect and significance (in the same genetic model and same direction) for the same SNP and the same allele as the initial report.[28] Replication is usually first attempted in studies as similar as possible to the initial report, but then may be extended to related phenotypes (such as fat mass in addition to obesity[44]), different populations (such as West Africans in addition to Icelanders[65]), or different study designs[33] to refine and extend the initial findings and increase confidence in verity.

Lack of reproducibility of genetic associations has been frequently observed and has been varyingly attributed to population stratification, phenotype differences, selection biases, genotyping errors, and other factors.[28,66] At present, the best way of resolving these inconsistencies appears to be additional replication studies with larger sample sizes, although this may not be feasible for rare conditions or for associations identified in unique populations.[28]

Identification of a robustly replicating SNP-disease association is a crucial first step in identifying disease-causing genetic variants and developing suitable treatments, but it is only a first step. Association studies essentially identify a genomic location related to disease but provide little information on gene function unless SNPs with predictable effects on gene expression or the transcribed product happened to be identified. Few of the associations identified to date have involved genes previously suspected of being related to the disease under study, and some have been in genomic locations harboring no known genes.[27,67] Examination of known SNPs in high linkage disequilibrium with the associated SNP may identify variants with plausible biologic effects, or sequencing of a suitable surrounding interval may be undertaken to identify rarer variants with more obvious functional implications. Tissue samples or cell lines can be examined for expression of the gene variant. Other functional studies may include genetic manipulations in cell or animal models, such as knockouts or knock-ins.[68]

Limitations of GWA Studies

The potential for false-positive results, lack of information on gene function, insensitivity to rare variants and structural variants, requirement for large sample sizes, and possible biases due to case and control selection and genotyping errors, are important limitations of GWA studies. The often limited information available about environmental exposures and other non-genetic risk factors in GWA studies will make it difficult to identify gene-environment interactions or modification of gene-disease associations in the presence of environmental factors. Clinicians and scientists should understand the unique aspects of these studies and be able to assess and interpret GWA results for themselves and their patients. Ten basic questions to ask about GWA studies, many of which also apply generically to association studies of nongenetic risk factors, are outlined in BOX 2. Most of these questions should be answered in the affirmative for a reliable report; however, many GWA reports lack sufficient detail to assess them.[28]

Many of the design and analysis features of GWA studies deal with minimizing the false-positive rates while maintaining powcr to identify true-positive associations. These same efforts to reduce false-positive results, however, may result in overlooking a true association, especially if only a small number of SNPs are carried over from the initial scan into replication studies. The most robust findings, ie, those that "survive" multiple rounds of replication, are often not the most statistically significant associations in the initial scan, and may not even be in the top few hundred associations.[69,70] Another cause of false-negative results is the lack of the genetic variant of relevance on the genotyping platform, or lack of variation in that SNP in the population under study. As the number of SNPs and diversity of populations represented on genotyping platforms increase, this should become less of a problcm.

An important question generated by these early GWA studies relates to the small proportion of heritability, or familial clustering explained by the genetic variants identified to date. Most of these variants have very modest effects on disease risk, increasing it by only 20% to 50%, and explaining only a small fraction of population risk or total estimated heritability for most conditions.[39,71] Might the rest of the genetic influence reside in a long "tail" of common SNPs with very small odds ratios, in copy number variants or other structural variants, rarer variants of larger effect, or interactions among common variants? Or has familial clustering due to genetic factors been overestimated and important environmental influences, either acting alone or in combination with genetic variants, been overlooked? This remains to be determined, but it is important to realize that even small odds ratios or rare variants can suggest important therapeutic strategies such as the development of HMG-CoA reductase inhibitors arising from identification of LDL-receptor mutations in familial hypercholesterolemia.[72]

Clinical Applications of GWA Findings

Despite the considerable media attention that GWA reports frequently receive, these studies are clearly many steps removed from actual clinical application. The primary use for GWA studies for the foreseeable future is likely to be in investigation of biologic pathways of disease causation and normal health and development. This is not to suggest that some early successes may not occur in the near future, through rapid development of treatment strategies such as inhibitors of complement activation in age-related macular degeneration.[73] Use of GWA findings in screening for disease risk, while beginning to be marketed commercially, is more problematic. Although obtaining the latest "gene test" may be alluring to a technology-focused society, evidence is needed that such screening adds information to known risk factors (such as age, obesity, and family history for diabetes), that effective interventions are available, that improved outcomes justify the associated costs, and that obtaining this information does not have serious adverse consequences for patients and their families. Such evidence is likely to be some ways off, but the initial burst of discovery generated by GWA scans has now mandated a concerted effort to search for these answers.

Author Contributions: Dr Manolio had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.
Financial Disclosures: None reported.
Funding/Support: This work supported in part by a Clinical and Translational Science Award (RR024160) from the National Center for Research Resources, National Institutes of Health, to the University of Rochester.
Additional Contributions: The authors wish to acknowledge the valuable comments of Francis Collins, MD, PhD, National Human Genome Research Institute, National Institutes of Health; and Stephen Chanock, MD, National Cancer Institute, National Institutes of Health; and the contributions of Maureen Marcello, Clinical and Translational Science Institute, University of Rochester, in the preparation of this article. None of these individuals received compensation for their work in association with this article.

REFERENCES

1. Christensen K, Murray JC. What genome-wide association studies can do for medicine. *N Engl J Med*. 2007;356(11):1094-1097.
2. National Center for Biotechnology Information, National Library of Medicine. Database of Single Nucleotide Polymorphisms. http://www.ncbi.nlm.nih.gov/SNP/. Accessed February 14, 2008.
3. Wheeler DL, Barrett T, Benson DA, et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. 2008;36(database issue):D13-D21.
4. Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet*. 2005;6(2):95-108.
5. Hunter DJ, Kraft P. Drinking from the fire hose—statistical issues in genomewide association studies. *N Engl J Med*. 2007;357(5):436-439.
6. Tabor HK, Risch NJ, Myers RM. Candidate-gene approaches for studying complex genetic traits. *Nat Rev Genet*. 2002;3(5):391-397.
7. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. A comprehensive review of genetic association studies. *Genet Med*. 2002;4(2):45-61.
8. Morgan TM, Krumholz HM, Lifton RP, Spertus JA. Nonvalidation of reported genetic risk factors for acute coronary syndrome in a large-scale replication study. *JAMA*. 2007;297(14):1551-1561.
9. Todd JA. Statistical false positive or true disease pathway? *Nat Genet*. 2006;38(7):731-733.
10. Patterson M, Cardon L. Replication publication. *PLoS Biol*. 2005;3(9):e327.
11. Altmüller J, Palmer LJ, Fischer G, Scherb H, Wjst M. Genome-wide scans of complex human diseases. *Am J Hum Genet*. 2001;69(5):936-950.
12. International HapMap Consortium. A haplotype map of the human genome. *Nature*. 2005;437(7063):1299-1320.
13. Frazer KA, Ballinger DG, Cox DR, et al; International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature*. 2007;449(7164):851-861.
14. Collins FS, Guyer MS, Chakravarti A. Variations on a theme: cataloging human DNA sequence variation. *Science*. 1997;278(5343):1580-1581.

15. National Institutes of Health. Policy for sharing of data obtained in NIH supported or conducted genome-wide association studies (GWAS) Federal Register. 2007;72(166):49290-49297. http://www.grants.nih.gov/grants/guide/notice-files/NOT-OD-07-088.html. Accessed February 14, 2007.
16. Hampe J, Franke A, Rosenstiel P, et al. A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. *Nat Genet.* 2007;39(2):207-211.
17. Weedon MN, Lettre G, Freathy RM, et al. A common variant of HMGA2 is associated with adult and childhood height in the general population. *Nat Genet.* 2007;39(10):1245-1250.
18. Arking DE, Pfeufer A, Post W, et al. A common genetic variant in the NOS1 regulator NOS1AP modulates cardiac repolarization (QT interval). *Nat Genet.* 2006;38(6):644-651.
19. Coon KD, Myers AJ, Craig DW, et al. A high-density whole-genome association study reveals that APOE is the major susceptibility gene for sporadic late-onset Alzheimer's disease. *J Clin Psychiatry.* 2007; 68(4):613-618.
20. Rioux JD, Xavier RJ, Taylor KD, et al. Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nat Genet.* 2007;39(5):596-604.
21. Reiman EM, Webster JA, Myers AJ, et al. GAB2 alleles modify Alzheimer's risk in APOE ε4 carriers. *Neuron.* 2001;54(5):713-720.
22. Thorleifsson G, Magnusson KP, Sulem P, et al. Common sequence variants in the LOXL1 gene confer susceptibility to exfoliation glaucoma. *Science.* 2007; 317(5843):1397-1400.
23. Gudbjartsson DF, Arnar DO, Helgadottir A, et al. Variants conferring risk of atrial fibrillation on chromosome 4q25. *Nature.* 2007;448(7151):353-357.
24. Moffatt MF, Kabesch M, Liang L, et al. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature.* 2007;448(7152): 470-473.
25. Dixon AL, Liang L, Moffatt MF, et al. A genome-wide association study of global gene expression. *Nat Genet.* 2007;39(10):1202-1207.
26. Manolio TA, Bailey-Wilson JE, Collins FS. Genes, environment and the value of prospective cohort studies. *Nat Rev Genet.* 2006;7(10):812-820.
27. Libioulle C, Louis E, Hansoul S, et al. Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. *PLoS Genet.* 2007;3(4):e58.
28. Chanock SJ, Manolio T, Boehnke M, et al; NCI-NHGRI Working Group on Replication in Association Studies. Replicating genotype-phenotype associations. *Nature.* 2007;447(7145):655-660.
29. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium. *Am J Hum Genet.* 1993;52(3):506-516.
30. Cardon LR, Palmer LJ. Population stratification and spurious allelic association. *Lancet.* 2003;361(9357): 598-604.
31. Mitchell AA, Cutler DJ, Chakravarti A. Undetected genotyping errors cause apparent overtransmission of common alleles in the transmission/disequilibrium test. *Am J Hum Genet.* 2003;72(3):598-610.
32. Cupples LA, Arruda HT, Benjamin EJ, et al. The Framingham Heart Study 100K SNP genome-wide association study resources. *BMC Med Genet.* 2007; 8(suppl 1):S1.
33. Ridker PM, Chasman DI, Zee RY, et al. Rationale, design, and methodology of the Women's Genome Health Study. *Clin Chem.* 2008;54(2):249-255.
34. Easton DF, Pooley KA, Dunning AM, et al. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature.* 2007;447(7148): 1087-1093.
35. Zanke BW, Greenwood CM, Rangrej J, et al. Genome-wide association scan identifies a colorectal cancer susceptibility locus or a chromosome 8q24. *Nat Genet.* 2007;39(8):989-994.
36. Skol AD, Scott LJ, Abecasis GR, Boehnke M. Optimal designs for two-stage genome-wide association studies. *Genet Epidemiol.* 2007;31(7):776-788.
37. Hoover RN. The evolution of epidemiologic research. *Epidemiology.* 2007;18(1):13-17.
38. Mueller PW, Rogus JJ, Cleary PA, et al. Genetics of Kidneys in Diabetes (GoKinD) study. *J Am Soc Nephrol.* 2006;17(7):1782-1790.
39. Wellcome Trust Case Control Consortium. Genome-wide association study of 14 000 cases of seven common diseases and 3 000 shared controls. *Nature.* 2007;447(7145):661-678.
40. Hakonarson H, Grant SF, Bradfield JP, et al. A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene. *Nature.* 2007;448(7153): 591-594.
41. Samani NJ, Erdmann J, Hall AS, et al; WTCCC and the Cardiogenics Consortium. Genomewide association analysis of coronary artery disease. *N Engl J Med.* 2007;357(5):443-453.
42. Scott LJ, Mohlke KL, Bonnycastle LL, et al. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science.* 2007; 316(5829):1341-1345.
43. Dewan A, Liu M, Hartman S, et al. HTRA1 promoter polymorphism in wet age-related macular degeneration. *Science.* 2006;314(5801):989-992.
44. Frayling TM, Timpson NJ, Weedon MN, et al. A common variant in the *FTO* gene is associated with body mass index and predisposes to childhood and adult obesity. *Science.* 2007;316(5826):889-894.
45. Thomas DC, Witte JS. Point: population stratification: a problem for case-control studies of candidate-gene associations? *Cancer Epidemiol Biomarkers Prev.* 2002;11(6):505-512.
46. Wacholder S, Rothman N, Caporaso N. Counterpoint: bias from population stratification is not a major threat to the validity of conclusions from epidemiological studies of common polymorphisms and cancer. *Cancer Epidemiol Biomarkers Prev.* 2002; 11(6):513-520.
47. Plenge RM, Seielstad M, Padyukov L, et al. TRAF1–C5 as a risk locus for rheumatoid arthritis—a genome-wide study. *N Engl J Med.* 2007;357(12):1199-1209.
48. Stacey SN, Manolescu A, Sulem P, et al. Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat Genet.* 2007;39(7):865-869.
49. Devlin B, Roeder K. Genomic control for association studies. *Biometrics.* 1999;55(4):997-1004.
50. Gabriel SB, Schaffner SF, Nguyen H, et al. The structure of haplotype blocks in the human genome. *Science.* 2002;296(5576):2225-2229.
51. McCarroll SA, Altshuler DM. Copy-number variation and association studies of human disease. *Nat Genet.* 2007;39(7)(suppl):S37-S42.
52. Feuk L, Marshall CR, Wintle RF, Scherer SW. Structural variants. *Hum Mol Genet.* 2006;15(Spec no 1): R57-R66.
53. Duerr RH, Taylor KD, Brant SR, et al. A genome-wide association study identifies *IL23R* as an inflammatory bowel disease gene. *Science.* 2006;314(5804): 1461-1463.
54. Moskvina V, Craddock N, Holmans P, et al. Effects of differential genotyping error rate on the type I error probability of case-control studies. *Hum Hered.* 2006;61(1):55-64.
55. Sladek R, Rocheleau G, Rung J, et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature.* 2007;445(7130):881-885.
56. Tomlinson I, Webb E, Carvajal-Carmona L, et al. A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q2421. *Nat Genet.* 2007;39(8):984-988.
57. Last JM, ed. *Dictionary of Epidemiology.* New York, NY: Oxford University Press; 1983:7.
58. Gordis L, ed. *Epidemiology.* 2nd ed. Philadelphia, PA: WB Saunders Co; 2000:165.
59. Zollner S, Pritchard JK. Overcoming the winner's curse: estimating penetrance parameters from case-control data. *Am J Hum Genet.* 2007;80(4):605-615.
60. Yang Q, Cui J, Chazaro I, Cupples LA, Demissie S. Power and type I error rate of false discovery rate approaches in genome-wide association studies. *BMC Genet.* 2005;6(suppl 1):S134.
61. Hochberg Y, Benjamini Y. More powerful procedures for multiple significance testing. *Stat Med.* 1990; 9(7):811-818.
62. Sabatti C, Service S, Freimer N. False discovery rate in linkage and association genome screens for complex disorders. *Genetics.* 2003;164(2):829-833.
63. Wacholder S, Chanock S, Garcia-Closas M, El Ghormli L, Rothman N. Assessing the probability that a positive report is false. *J Natl Cancer Inst.* 2004; 96(6):434-442.
64. Todd JA, Walker NM, Cooper JD, et al. Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nat Genet.* 2007;39(7):857-864.
65. Helgason A, Pálsson S, Thorleifsson G, et al. Refining the impact of TCF7L2 gene variants on type 2 diabetes and adaptive evolution. *Nat Genet.* 2007; 39(2):218-225.
66. Khoury MJ, Little J, Gwinn M, Ioannidis JPA. On the synthesis and interpretation of consistent but weak gene-disease associations in the era of genome-wide association studies. *Int J Epidemiol.* 2007;36(2):439-445.
67. Yeager M, Orr N, Hayes RB, et al. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat Genet.* 2007;39(5):645-649.
68. Frayling TM, McCarthy MI. Genetic studies of diabetes following the advent of the genome-wide association study: where do we go from here? *Diabetologia.* 2007;50(11):2229-2233.
69. Hunter DJ, Kraft P, Jacobs KB, et al. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat Genet.* 2007;39(7):870-874.
70. Hafler DA, Compston A, Sawcer S, et al; The International Multiple Sclerosis Genetics Consortium. Risk alleles for multiple sclerosis identified by a genome-wide study. *N Engl J Med.* 2007;357(9):851-862.
71. Saxena R, Voight BF, Lyssenko V, et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science.* 2007;316(5829): 1331-1336.
72. Slater EE, MacDonald IS. Mechanism of action and biological profile of HMG CoA reductase inhibitors: a new therapeutic alternative. *Drugs.* 1988;36(suppl 3): 72-82.
73. Gehrs KM, Anderson DH, Johnson LV, Hageman GS. Age-related macular degeneration—emerging pathogenetic and therapeutic concepts. *Ann Med.* 2006;38(7):450-471.

conduct such studies are likely to assume that their work falls under this category. Conversely, public health and health services researchers might be confused to see the term applied to their work.

Furthermore, using "clinical" in both terms perpetuates the tendency of the medical profession to view health research through the clinician's lens alone. Fiscella et al do include "organizational- and community-focused" research within their definition of applied clinical research, but labeling health interventions outside the clinic as "clinical" research may be a forced fit. Pros and cons exist with other potential terms such as *knowledge translation*—the term discussed by Dr Graham and Ms Tetroe—but all of them are an improvement over the ambiguity of T2.

Graham and Tetroe call attention to the excellent work of the Canadian Institutes of Health Research. Canadian investigators and institutions have played a leadership role not only in writing about the need for researchers to align their work with the information needs of end users[1] but also in making real commitments in programs and funding to facilitate T2 as a nation.[2] The United States would do well to follow the Canadian example.

T1 is among a group of clinical research movements that are attracting attention and resources but are ultimately unhelpful to patients without T2. Recently, politicians and industry have announced plans to channel millions of dollars per year into research on "comparative effectiveness"[3] and "personalized medicine"[4] while keeping funding for health services research threadbare.[5] Popular research initiatives address worthy questions: whether a treatment can be produced (T1), whether it improves health (evidence-based medicine), which treatment is best (comparative effectiveness), and which is best for an individual patient (personalized medicine). But the answers remain academic if the patient cannot obtain or use the intervention. Overcoming such obstacles so that the products of research benefit all those in need is itself a crucial research priority.

Steven H. Woolf, MD, MPH
swoolf@vcu.edu
Department of Family Medicine
Virginia Commonwealth University
Richmond

Financial Disclosures: None reported.

1. Ross S, Lavis J, Rodriguez C, Woodside J, Denis JL. Partnership experiences: involving decision-makers in the research process. *J Health Serv Res Policy*. 2003; 8(suppl 2):26-34.
2. Lomas J. Using 'linkage and exchange' to move research into policy at a Canadian foundation. *Health Aff (Millwood)*. 2000;19(3):236-240.
3. US House of Representatives Committee on Ways and Means Subcommittee on Health: hearing on strategies to increase information on comparative clinical effectiveness [Tuesday, June 12, 2007]. http://waysandmeans.house.gov/hearings.asp?formmode=detail&hearing=565. Accessed February 29, 2008.
4. US Department of Health and Human Services: remarks prepared for the Honorable Mike Leavitt, Secretary of Health and Human Services, on personalized medicine coalition. http://www.hhs.gov/news/speech/2007/sp20070919a.html. Accessed February 29, 2008.
5. Coalition for Health Services Research: federal funding for health services research (as of December 26, 2007). http://www.chsr.org/2008.pdf. Accessed February 29, 2008.

CORRECTIONS

Incorrect Data: In the Perspectives on Care at the Close of Life article titled "Managing an Acute Pain Crisis in a Patient With Advanced Cancer: 'This Is as Much of a Crisis as a Code' " published in the March 26, 2008, issue of *JAMA* (2008;299 (12):1457-1467), an incorrect dose ratio appeared in Table 2. The hydromorphone-to-methadone ratio for less than 330 mg/24 hours of hydromorphone that read "16:1" should have been "1.6:1."

Incorrect Legend: In the Special Communication entitled "How to Interpret a Genome-wide Association Study" published in the March 19, 2008, issue of *JAMA* (2008;299(11):1335-1344), an integral word was omitted from the Figure 3 legend. The sentence that read, "Genome-wide association studies assume a priori hypotheses about candidate genes or regions that might be associated with disease; rather, they test single-nucleotide polymorphisms (SNPs) throughout the genome for possible evidence of genetic susceptibility" should have read, "Genome-wide association studies assume no a priori hypotheses about candidate genes or regions that might be associated with disease; rather, they test single-nucleotide polymorphisms (SNPs) throughout the genome for possible evidence of genetic susceptibility."

Unreported Research Funding: In the Research Letter titled "Exhaled Carbon Monoxide With Waterpipe Use in US Students," published in the January 2, 2008, issue of *JAMA* (2008;299(1):36-38), the Financial Disclosures should have included the following: Dr Hammond reports that she has received research funding for studies on environmental tobacco smoke from the National Institutes of Health and from the Flight Attendants Medical Research Institute. However, none of these grants were used to support the study reported in this Research Letter.

Bowcock, AM, Nature 447:645-46 (2007)

GENOMICS

Guilt by association

Anne M. Bowcock

In a tour-de-force demonstration of feasibility, a consortium of 50 research teams uses 500,000 genetic markers from each of 17,000 individuals to identify 24 genetic risk factors for 7 common human diseases.

Mr Woodhouse, the comical hypochondriac of Jane Austen's *Emma*, takes great comfort in blaming his various ailments on the rain, the cold and an unfortunate piece of wedding cake. He would, no doubt, have been greatly surprised to learn that even his most rudimentary ailments resulted, at least in part, from genetic factors. Reporting on page 661 of this issue[1], a consortium of more than 50 British groups, known collectively as the Wellcome Trust Case Control Consortium (WTCCC), asserts just that. In the largest study of its type so far, the WTCCC has examined the genetic underpinnings of seven common human diseases: rheumatoid arthritis, hypertension, Crohn's disease (the most common form of inflammatory bowel disease), coronary artery disease, bipolar disorder — also known as manic depression — and type 1 and type 2 diabetes.

The WTCCC study is groundbreaking in various respects. It not only confirms the involvement of some genes for which disease association has previously been reported, but it also identifies several novel genes that affect susceptibility to common diseases. Moreover, it models a successful and instructive approach to large-scale genomic scans of this type, showing that a set of common controls can be used for a variety of diseases with relatively little loss of analytical power. Its success also provides strong grounds for performing such studies on an even larger scale.

The WTCCC investigators examined genetic variation at 500,000 different positions within the genomes of 17,000 individuals living in Britain using a genome-wide association scan (Fig. 1). This statistical approach compares the frequencies of genetic variation in disease cases and in healthy controls from the same population. Using the signal from each position as an indicator for the DNA sequence that surrounds it, genome-wide association scans examine the relationship between each DNA position and a particular trait (such as diabetes). Strong 'association' between a DNA position and a trait marks the general locale of the offending alteration, even if it is not itself the cause.

The concept of drawing an association between biological traits and disease is hardly new[2], but the scope and scale that the WTCCC attained in their application of this concept is unprecedented. Crucial to both the success of this study and keeping its cost reasonable were DNA from large numbers of unrelated patients; the availability of the complete DNA sequence of the human genome; the subsequent cataloguing of a large component of variation in the genome in the form of single nucleotide polymorphisms (SNPs)[3]; the completion of the HapMap project[4], which provided information on the statistical relatedness of SNPs; and the availability of high-throughput technologies that allowed for parallel typing of 500,000 markers representing most of the common variation in the genome.

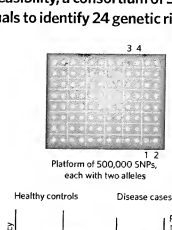


Figure 1 | Genome-wide association scan. To identify genetic risk factors for common diseases, the WTCCC researchers[1] scanned DNA from patients (2,000 per disease) and controls (3,000 shared for all seven diseases studied) for the frequency with which they contained each of the 500,000 genetic markers, or single nucleotide polymorphisms (SNPs), from the human genome. After statistical evaluation of the data, they found that most markers showed very little difference in the frequency of their two constituent forms — or alleles — between controls and cases. However, some SNPs occurred at a greater frequency in patients. Such alleles (one is shown in red) can be considered a genetic risk factor for a particular disease.

For the seven diseases studied by the WTCCC, strong statistical evidence for association was obtained for 12 previously identified genomic regions and a similar number of new regions. Although this WTCCC report is based on initial studies, independent groups[5–9] have confirmed the involvement of all but one of these most significant regions through replication studies. Some of the other identified regions with less statistically significant disease association are also likely to be true indicators of genetic risk; so these will need to be further evaluated in additional large sets of patients and controls. Indeed, because the WTCCC data will be publicly available, they will be a useful resource to other groups and consortia embarking on similar efforts to investigate genetic-association markers in these and other diseases. These researchers include members of the Genetic Association Information Network[10] (GAIN), the Framingham Genetic Research Study and the Women's Health Study.

With many of the genomic regions identified by the WTCCC, the next step will be to study the exact nature of the disease-causing variants, rather than the marker SNP with which each is associated. From this and previous studies, it seems that variations leading to common disease are diverse; some alter the coding sequences of genes, others lie within their non-coding sequences, and some are even located within gene deserts — regions of a chromosome that contain no genes. So understanding the biological function of disease-risk-associated genomic regions will be challenging.

Two replication studies relating to the WTCCC findings are also published today[5,6], revealing connections between the genomic regions associated with the risk of type 1 diabetes and Crohn's disease and their underlying biology. Some of the known and newly identified genetic risk factors for type 1 diabetes alter the development or function of immune cells, leading to aberrant recognition of pancreatic islet cells as foreign particles. But additional susceptibility genes identified recently[5] do not fit easily into this simple model.

For Crohn's disease, one of the newly identified[6] susceptibility genes is of particular interest because it is proposed to control the spread

NATURE Vol 447 7 June 2007

of intracellular pathogens by autophagy — the process of cellular self-digestion. This is the second gene to be implicated in Crohn's disease through involvement in autophagy; the first was identified earlier this year[11,12]. Moreover, an increasing body of evidence, including the latest replication study[6], points to defects in the early immune response and the handling of intracellular gut bacteria in the pathogenesis of Crohn's disease.

The overall increase in risk (1.2–1.5 times) conferred by the genetic factors identified in the WTCCC study[1] is in agreement with those reported by others. However, these factors are unlikely to explain completely the clustering of any of these diseases in families, and there are other genes (possibly many of very small effect) — or rare variants of genes — that are still to be identified for these and other diseases.

One unexpected result of the WTCCC study was the identification of 13 regions with pronounced geographical variation within Britain. Among these regions is a large cluster of genes that encodes the major histocompatibility complex, which is well known for its function in the immune response and autoimmune disease[13], and a gene that is involved in lactase persistence, or the ability to digest milk[14,15]. Some of the other regions are thought to function in preventing diseases such as pellagra, tuberculosis and leprosy. Although the infectious agents responsible for tuberculosis and leprosy are now rare in Britain, they have left behind genetic footprints in the existing population that probably led to some degree of protection in the past. Several of these are also candidate genes for autoimmune disease[5].

Despite the magnitude and wealth of information that this study[1] provides, other questions about the genetic basis of common disease remain. The answers will become increasingly important as we enter an era of personalized medicine, in which therapy is tailored to an individual's genetic constitution. It will become crucial to discover which genes predispose individuals to these diseases; how genes interact with each other to increase the risk of a particular disease; and what proportion of disease is due to rare variants that would be hard to detect with current approaches.

We will also want to know whether different patients can be stratified into subpopulations on the basis of genetic risk factors, and what role the environment has in triggering disease. The Genes, Environment and Health Initiative (GEI) of the US National Institutes of Health already aims to develop tools to assess environmental contribution and to answer some of the other questions. Ultimately, comprehensive answers that would allow the translation of genetic susceptibility into scientifically sound medical practice will require much larger patient populations, well-annotated clinical databases and sophisticated environmental assessment. One wonders what Mr Woodhouse would have to say to that. ■

Anne M. Bowcock is in the Departments of Genetics, Pediatrics and Medicine, Division of Human Genetics, Washington University School of Medicine, 4559 Scott Avenue, Saint Louis, Missouri 63110, USA.
e-mail:bowcock@genetics.wustl.edu

1. Wellcome Trust Case Control Consortium *Nature* **447**, 661–678 (2007).
2. Buckwalter, J. A., Wohlwend, C. B., Colter, D. C., Tidrick, R. T. & Knowler, L. A. *Surg. Gynecol. Obstet* **104**, 176–179 (1957).
3. Carlson, C. S. *et al. Am. J. Hum. Genet.* **74**, 106–120 (2004).
4. International HapMap Consortium *Nature* **437**, 1299–1320 (2005).
5. Todd, J. A. *et al. Nature Genet.* doi:10.1038/ng2068 (2007).
6. Parkes, M. *et al. Nature Genet.* doi:10.1038/ng2061 (2007).
7. Zeggini, E. *et al. Science* doi:10.1126/science.1142364 (2007).
8. Saxena, R. *et al. Science* doi:10.1126/science.1142358 (2007).
9. Frayling, T. M. *et al. Science* doi:10.1126/science.1141634 (2007).
10. www.hnih.org/GAIN2/home_new.shtml
11. Hampe, J. *et al. Nature Genet.* **39**, 207–211 (2007).
12. Rioux, J. D. *et al. Nature Genet.* **39**, 596–604 (2007).
13. Tomlinson, I. P. & Bodmer, W. F. *Trends Genet.* **11**, 493–498 (1995).
14. Cavalli-Sforza, L. *Am. J. Hum. Genet.* **25**, 82–104 (1973).
15. Enattah, N. S. *et al. Nature Genet.* **30**, 233–237 (2002).

SPECTROSCOPY

The magic of solenoids

Arthur S. Edison and Joanna R. Long

A technique known as magic-angle spinning has helped make nuclear magnetic resonance spectroscopy as sensitive for solids as it is for solutions. Inductive thinking leads to even better signal detection.

The great strength of nuclear magnetic resonance (NMR) spectroscopy is that it can determine, non-invasively and at atomic resolution, the chemistry, structure, dynamics and overall architecture of samples in solid, liquid or even gaseous forms. The liquid version of the technique, solution NMR, is used routinely to identify small molecules, study protein structures and dynamics, and probe intermolecular interactions. Solid-state NMR teases out the structure and properties of materials, surfaces and biological solids such as human tissue. But compared with many other analytical techniques, NMR has extremely poor sensitivity. A great deal of research has sought to improve this situation: on page 694 of this issue[1], Sakellariou *et al.* describe a potential leap forward for solid-state NMR.

When atomic nuclei with non-zero spin are placed in an external magnetic field, they become polarized, precessing rather as a gyroscope does in Earth's gravitational field. When electromagnetic radiation of a frequency (energy) that corresponds exactly to that of the energy gap between two states of different polarization is applied to the sample, the nuclei resonate, jumping between those states. The accompanying gyroscopic precession of the spins induces a current in a conducting coil placed around the sample. This basic principle is both NMR's blessing and its bane as a spectroscopic technique: the small energies make the approach non-destructive, but they also make it difficult to distinguish the characteristic polarization (or signal) from thermal noise.

The signal-to-noise ratio in NMR measurements can be improved by either one of two general routes. The first of these is enhancing

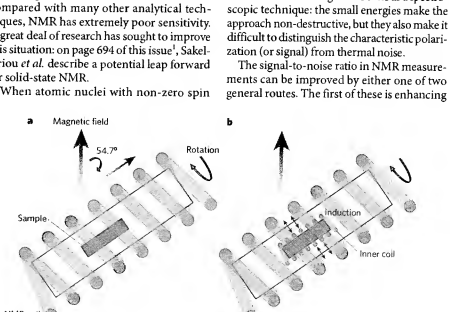


Figure 1 | Inductive logic. a, In the traditional 'magic-angle spinning' approach to solid-state NMR, a spectrum of better resolution is achieved by rapidly rotating the sample, at an angle of 54.7° relative to the main magnetic field, within a static coil assembly. **b**, Sakellariou and colleagues' alternative approach[1] uses the inductive coupling of a smaller coil rotating with the sample to the larger static coil to produce a similar effect. The result is a higher sensitivity and the capability to investigate smaller samples.

Altshuler, D & Daly, M., Nature Genet 39:813-815 (2007)

Guilt beyond a reasonable doubt

David Altshuler & Mark Daly

Genome-wide association studies, exemplified by the Wellcome Trust Case Control Consortium and follow-up studies, have identified dozens of common variants robustly associated with common diseases, providing new clues about genetic architecture in humans. Finding all such loci, and fully defining genotype-phenotype correlation, will be a key to translating initial clues into pathophysiological understanding and clinical prediction.

Genetic screens are used to explore biological mechanisms *in vivo*, unbiased by prior assumptions about the DNA alterations responsible for phenotypic variation. In model systems, genome-wide, phenotype-driven screens typically identify many genes of unknown function, ultimately leading to a broad and deep understanding of mechanism.

In humans, success with phenotype-driven, genome-wide screening for inherited disease mutations has been limited to mendelian traits. Human phenotypic variation is largely polygenic rather than monogenic, however, and thus the vast majority of heritable factors for common human diseases remain unknown. Genome-wide association studies (GWASs) have been proposed as a new approach to 'forward genetics' in humans, but until recently they were untested for gene discovery.

The Wellcome Trust Case Control Consortium (WTCCC) now reports in *Nature* the largest GWAS thus far[1], scanning 17,000 individuals for seven diseases, with two follow-up studies reported in this issue, Todd *et al.* on type 1 diabetes (page 857) and Parkes *et al.* on Crohn's disease (page 830), and another on type 2 diabetes published elsewhere[2–4]. Together with other publications, statistically compelling associations have been identified this year by GWASs across a variety of diseases, including Crohn's disease, obesity, type 1 and type 2 diabetes, coronary heart disease and prostate and breast cancer (see **Supplementary Note** for additional references). In multiple diseases, five to ten independent genomic regions have been identified and confirmed. After years as 'Keystone Cops', complex trait geneticists can now find culprits not previously suspected and establish guilt beyond a reasonable doubt.

The current crop of successful studies shares five key features. First, they all use high-density SNP genotyping arrays (based on the Human Genome Project, the SNP Consortium and the HapMap Project) and analytical methods built on the synthesis of population genetics, statistical genetics and epidemiology. Second, the clinical investigators had the foresight to collect large patient samples that included detailed phenotype information, DNA samples and informed consent for genetic research. Third, they have paid careful attention in their design and analysis to minimizing bias (coming from, for example, population substructure, genotyping errors or variability in DNA quality and laboratory processing). Fourth, they have applied statistical thresholds appropriate to genome-wide searches. With ~10 million common SNPs to be tested genome-wide, and few true associations for which power is adequate, the prior probability of a true association is low—and the *P* value required to declare significance is correspondingly stringent (for further discussion of power in the WTCCC, see pages 815–816 in this issue). Finally, they have validated putative 'positives' in independent samples (preferably using independent genotyping technologies). Here, 'replication' refers to association of the same allele to the same trait under the same genetic model[5].

What has been learned?

The most important outcome of these studies is the discovery of new biological associations in genes or regions previously unrecognized to have a role in each disease. In some cases, links have been newly established between diseases and well-studied pathways (such as age-related macular degeneration and the complement pathway, Crohn's disease and autophagy). In many cases, however, associated regions contain genes of unknown function or do not contain annotated genes. Typical of genetic screens in model systems and mendelian genetics, an unbiased genetic approach highlights genes not previously identified.

Second, new mechanistic connections have been uncovered between diseases. Examples include SNPs in *IL23R* with Crohn's disease[6] and psoriasis[7], *PTPN2* with Crohn's disease and type 1 diabetes[1], *PTPN22* and *IL2RA* with type 1 diabetes and rheumatoid arthritis[1], 8q24 with prostate cancer and breast cancer[8] (see also Stacey *et al.* (page 865) and Hunter *et al.* (page 870), in this issue) and nearby SNPs in a noncoding region of 9p near *CDKN2B* and *CDKN2A* with type 2 diabetes[4,9,10] and coronary heart disease[1,11,12].

Third, the studies have found a substantial fraction of associations outside of transcription units. This is unsurprising, as coding sequences make up less than half of the evolutionarily conserved DNA in the human genome. Investigation of functional noncoding associations will be critical to unraveling molecular and cellular roles of noncoding functional DNA in humans.

David Altshuler and Mark Daly are at Massachusetts General Hospital, Harvard Medical School and the Broad Institute of Harvard and MIT, Boston, Massachusetts, USA. e-mail: altshuler@molbio.mgh.harvard.edu and mjdaly@chgr.mgh.harvard.edu

Table 1 Power of GWASs to discover several recently defined associations

		Power in a 'typical' GWAS (1,000 cases/1,000 controls)			Power in WTCCC (2,000 cases/3,000 controls)			Sample size required for 90% power,		
Gene	Disease	1.0×10^{-2}	1.0×10^{-4}	1.0×10^{-6}	1.0×10^{-2}	1.0×10^{-4}	1.0×10^{-6}	$P < 10^{-8}$	RAF	RR
ATG16L1	CD	>0.99	>0.99	0.74	>0.99	>0.99	>0.99	2,430	0.5	1.5
IRGM	CD	0.67	0.19	<0.01	0.98	0.8	0.16	10,902	0.075	1.4
PTPN2	T1D, CD	0.37	0.05	<0.01	0.82	0.34	<0.01	19,754	0.17	1.2
IL2	T1D	0.11	<0.01	<0.01	0.31	0.04	<0.01	54,600	0.26	1.1
9p21	MI	0.97	0.87	0.09	>0.99	>0.99	0.86	5,066	0.47	1.25
9p21	T2D	0.36	0.05	<0.01	0.79	0.31	<0.01	20,220	0.83	1.2
CDKAL1	T2D	0.35	0.04	<0.01	0.79	0.31	<0.01	20,700	0.31	1.15

Approximate risk models estimated from published replication studies and power computed using the Genetic Power Calculator[15] (http://pngu.mgh.harvard.edu/~purcell/gpc/). Sample size calculation assumes equal numbers of cases and controls. RAF, risk allele frequency; RR, relative risk; CD, Crohn's disease; T1D, type 1 diabetes; MI, myocardial infarction; T2D, type 2 diabetes; WTCCC, Wellcome Trust Case Control Consortium.

Fourth, the results indicate that individual SNPs have very modest effects in the population: associated SNPs rarely show odds ratios of >2.0 (CFH in age-related macular degeneration), and more typically, odds ratios are <1.5. Undiscovered common variants are likely to have similar or smaller effects (or are in low linkage disequilibrium with SNPs on arrays).

Fifth, strong evidence is lacking for epistasis among associated SNPs, despite joint analysis in large cohorts. Similarly, little evidence has been obtained for strong association of disease-associated SNPs to homogeneous disease subtypes, or quantitative 'endo-phenotypes' (such as glycemic and obesity traits in type 2 diabetes). Sixth, despite substantial progress, the vast majority of heritability remains unexplained. To some extent, the magnitude of the associations discovered is currently underestimated, because the full spectrum of causal variation at each locus has yet to be defined by deep sequencing.

A less obvious but still important implication is that many more such loci must remain to be found. Even for the confirmed associations identified, statistical power was limited in the genome-wide scans that found them (Table 1). Even in the large WTCCC study (which included 2,000 cases and 3,000 controls)[1], the power to obtain a genome-wide $P < 10^{-8}$ was <1% for many of the confirmed associations discovered by comparison across studies and by replication studies. This explains the tendency of different GWASs to find partially overlapping sets of associations and makes it implausible that most regions harboring relevant associations have been identified.

Where to from here?

These papers provide proof-of-concept that GWASs can identify previously unknown causal loci. The next steps are to obtain a full picture of genotype-phenotype correlation at these loci and to find remaining loci. A more complete picture will be critical to understanding the disease mechanisms underlying the associations and to assess SNPs for clinical management.

Rarely will the SNPs used to discover each locus prove causal; exhaustive sequencing of each region will be needed to discover all causal mutations and fully define genotype-phenotype correlation. In many cases, multiple independent common variants[13] and rare variants[14] will be found at the same locus. Sequencing of exons in each associated region may identify coding mutations of stronger effect, which may be easier to study *in vitro* and in individual subjects. In addition, identification of 'smoking gun' causal coding mutations may help prove which gene at each locus is responsible for the association and may, in aggregate, increase the overall predictive value of genotype.

A testable hypothesis suggested by the power calculations in **Table 1** is that a more extensive set of loci that influence each disease may be found by GWASs of greater power (or by combining existing GWASs). Common sense dictates that a complete set of susceptibility loci will provide greater biological insight than an incomplete set. Moreover, the biological insight provided by any locus is not necessarily related to the size of the effect of common variants used to discover it, nor is it predictive of the combined effect of all rare and common variants at that locus. Thus, the discovery of additional causal loci should be pursued, followed by exhaustive sequencing to fully define genotype-phenotype correlation.

Some loci may be missed by well-powered GWASs because none of the causal variants are in linkage disequilibrium with SNPs on the genotyping arrays. Some of these may be found by genome-wide measurement of copy number variation. Thus, these GWASs are the first in a series of genome-wide, phenotype-driven approaches in humans, which, when integrated, will provide a more complete picture of human phenotype variation and inborn susceptibility to disease.

Ultimately, the value of this endeavor must be measured in the resulting clinical and biological advances. Predictive testing will have value in cases in which effective preventative interventions exist, and when modest changes in risk improve clinical decision-making. Achieving a clinical benefit will be challenged by the modest magnitude of SNP effects and by the likelihood that genetic tests will be made available (and aggressively promoted) before or instead of mounting clinical trials to evaluate the value of genetically enabled decision-making.

New tools and frameworks will be required to translate genetic insights into knowledge of disease pathogenesis and new therapeutics: there is little precedent for functional analysis based on genes discovered by polygenic inheritance, noncoding DNA changes and quantitative alteration of gene function. This quest is worth mounting, however, as it is in pursuit of culprits whose guilt in human disease has been established beyond a reasonable doubt.

Note: Supplementary information is available on the Nature Genetics website.

COMPETING INTERESTS STATEMENT
The authors declare no competing financial interests.

1. Wellcome Trust Case Control Consortium. *Nature* **447**, 661–678 (2007).
2. Parkes, M. *et al. Nat. Genet.* **39**, 830–832 (2007).
3. Todd, J. *et al. Nat. Genet.* **39**, 857–864 (2007)
4. Zeggini, E. *et al. Science* **316**, 1336–1341 (2007).
5. NCI-NHGRI Working Group on Replication in Association Studies *et al. Nature* **447**, 655–660 (2007).
6. Duerr, R.H. *et al. Science* **314**, 1461–1463 (2006).
7. Cargill, M. *et al. Am. J. Hum. Genet.* **80**, 273–290 (2007).
8. Easton, D.F. *et al. Nature*, advance online publication 27 May 2007 (doi:10.1038/nature05887).
9. Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of BioMedical Research *et al. Science* **316**, 1331–1336 (2007).
10. Scott, L.J. *et al. Science* **316**, 1341–1345 (2007).

11. McPherson, R. *et al. Science*, **316**, 1488–1491 (2007).
12. Helgadottir, A. *et al. Science*, **316**, 1491–1493 (2007).
13. Haiman, C. *et al. Nat. Genet.* **39**, 638–644 (2007).
14. Kotowski, I.K. *et al. Am. J. Hum. Genet.* **78**, 410–422 (2006).
15. Purcell, S. *et al. Bioinformatics* **19**, 149–150 (2003).

Conjuring SNPs to detect associations

Andrew G Clark & Jian Li

Human genome-wide association studies pose a challenge in identifying significant disease associations from nearly half a million statistical tests. A new report describes an especially promising approach, recently applied to the Wellcome Trust Case Control Consortium data sets, that uses the correlated structure of genomic variation to impute genotypes at missing sites and to test association with both observed and imputed SNPs.

Genetic mapping has always relied on statistical inference, but this enterprise has never been so utterly dependent on rigorous analytical methods as it is with genome-wide association studies (GWASs). For each of the nearly 500,000 SNPs in the human genome scored by widely used genotyping platforms for GWASs, it is possible to perform a simple statistical test of association with disease state. Even if the null hypothesis of no association were true for all SNPs, we would expect some of these tests to provide nominal P values on the order of 10^{-6}. In order to avoid false-positive calls, we need to identify SNPs for which the P values are even lower. We could increase the power to appropriately reject the null hypothesis (that is, to correctly infer that a SNP is truly associated with disease) by elevating the sample size or restricting attention to intermediate-frequency SNPs and by being judicious in our choice of test. In this issue, Marchini *et al.*[1] (page 906) show that thoughtful application of population genetic principles and use of HapMap data can provide an additional source of power for association tests. They have successfully applied these methods to the Wellcome Trust Case Control Consortium (WTCCC) data[2] and have identified a collection of new genes associated with seven complex medical disorders (see pages 813–815 of this issue for discussion of the WTCCC studies).

Imputation to boost power

The more genetic data that we have for each individual, the greater the chance of finding variants that influence disease risk directly. This is true even if some of those variants are statistically inferred or 'imputed' from the observed genetic data. To see how imputation can give a boost in the power of tests of association, consider the situation where a SNP that has a direct effect on disease risk is in the HapMap set of SNPs but is not on the 500K genotyping platform used in a given GWAS (**Fig. 1a**). In this case, if only the observed marker SNP were used, the association test would be weakened by any observed departure from perfect linkage disequilibrium between the observed SNP and the unobserved risk-enhancing SNP. This contrasts with the hypothetical case (**Fig. 1b**) in which the risk-enhancing SNP is observed directly. If no other genetic variation in this genomic region influences risk, then the test based on this SNP alone will be the most powerful. One can see that such a direct test provides a greater chance to detect a significant association. Because we often do not observe the risk-enhancing SNP directly, imputation can be used to close some of the gap between these two extremes. High linkage disequilibrium in the human genome means that we can impute the unobserved genotype of many of the missing SNPs with surprisingly high accuracy (>98% in many cases). This accuracy will be reduced in regions of the genome with unusually high recombination rates (for example, SNPs within hotspots). The example in **Figure 1c** is for an imputation accuracy of 99%, and it is clear that the probability of detecting the association is much greater than in **Figure 1a**, where we did not apply imputation. Marchini *et al.*[1] and Scott *et al.*[3] use multiple flanking SNPs to impute missing SNP genotypes, and they find that the P values for tests of association are often an order of magnitude lower with the imputed SNPs than with the observed SNP data only.

This may seem like sleight of hand, because there seems to be a gain in power without any additional information, as the missing SNPs are imputed from the observed marker SNPs. One might think that tests based only on haplotypes of the observed SNPs[4–6] would do just as well, because they, after all, are what allows prediction of the missing SNPs. But the method does incorporate haplotype information of observed SNPs along with the linkage disequilibrium structure of the full HapMap sample to perform the imputation. By leveraging the observed marker SNPs and by predicting missing data from the pattern of linkage disequilibrium in the HapMap data, we get the best of both worlds.

Testing association

In a GWAS, the meaning of a P value becomes challenged in the context of so many simultaneous tests. One solution to this problem is to calculate the false discovery rate[7,8]; however, this approach was developed for testing a single hypothesis, as opposed to simultaneously testing a battery of SNPs associated with a disease. Association testing can be done with standard frequentist methods like logistic regression, where the model may specify either allelic or genotypic effects. Likelihood methods can be used to deal with the uncertainty in the imputations of missing genotype data. Bayesian methods also allow inference of probability of association conditional on observed genotype data and can accommodate imputed genotypes easily. Marchini *et al.*[1] make use of one useful measure of the relative likelihood of association, the Bayes factor, a term closely related to likelihood ratio and defined in this case as the probability of the observed data, given that the association is real, divided by the probability of the observed data under

Andrew G. Clark and Jian Li are in the Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York 14853, USA.
e-mail: ac347@cornell.edu

Kruglyak, L, Nature Reviews Genet 9:314-18 (2008)

ESSAY

The road to genome-wide association studies

Leonid Kruglyak

Abstract | The recent crop of results from genome-wide association studies might seem like a sudden development. However, this blooming follows a long germination period during which the necessary concepts, resources and techniques were developed and assembled. Here, I look back at how the necessary pieces fell into place, focusing on the less well-chronicled days before the launch of the HapMap project, and speculate about future developments.

Genome-wide association studies (GWAS) use dense maps of SNPs that cover the human genome to look for allele-frequency differences between cases (patients with a specific disease or individuals with a certain trait) and controls. A significant frequency difference is taken to indicate that the corresponding region of the genome contains functional DNA-sequence variants that influence the disease or trait in question. The recent crop of results from GWAS (reviewed in REFS 1–4) might seem like a sudden development. However, this blooming follows a long germination period during which the necessary concepts, resources and techniques were developed and assembled. Here, I look back at how the necessary pieces fell into place, starting with early ideas and continuing with concrete proposals and theoretical and empirical studies that laid the foundation for The International HapMap Project[5]. I close by contemplating the implications of the lessons that were learnt from the initial crop of GWAS for future studies of human genetic variation.

Early milestones

Genome-wide approaches to human genetics date back to the proposal in 1980 by Botstein and colleagues for the construction of a linkage map of the human genome, with restriction fragment length polymorphisms (RFLPs) as molecular markers[6,7]. The natural initial applications were to genetically simple Mendelian diseases; however, as early as 1986, even before the first linkage map was completed, Lander and Botstein recognized that most human traits and diseases follow complex modes of inheritance, and they discussed several approaches for studying such complex traits[8]. One of the approaches they proposed was linkage disequilibrium (LD) mapping, which recognizes that a mutation that is shared by affected individuals through common descent will be surrounded by shared alleles at nearby loci, representing the haplotype of the ancestral chromosome on which the mutation first occurred (FIG. 1). The first example of LD between a DNA polymorphism and a disease mutation was provided by an association between an allele of an RFLP in the β-globin gene and the sickle-cell form of haemoglobin[9].

> [Genetic] complexity is present on multiple levels, and might be fruitfully thought of as 'fractal'.

Simple population-genetics arguments suggested that LD in the general human population would probably be limited to distances below 100 kb[10]. For this reason, Lander and Botstein deemed LD mapping to be impractical in the general population owing to the high marker density that would be required, but they proposed that a map of hundreds of RFLPs might suffice for LD mapping in recently founded isolated populations.

The first complete RFLP map of the human genome was reported in 1987 (REF. 11), but human mapping studies really flourished once microsatellites replaced RFLPs[12] (BOX 1). Genome-wide studies used family-linkage approaches almost exclusively, with LD being used to refine the locations of genes that were mapped by linkage, as pioneered by Kerem and colleagues for the cystic fibrosis gene[13]. In a groundbreaking and forward-looking study in 1994, Houwen and colleagues reported the first application of LD mapping in an initial whole-genome search for a disease locus[14]. Following the approach that was envisioned by Lander and Botstein almost a decade earlier, they used 256 markers to map the gene responsible for benign recurrent intrahepatic cholestasis (BRIC) in an isolated fishing community in the Netherlands. Their success relied on the rarity of the disease and on the availability of a population isolate in which the affected individuals were distant relatives. A similar study in a Mennonite kindred allowed Puffenberger and colleagues to identify a gene for Hirschsprung disease[15]. However, such studies, which straddled the border between family linkage and LD mapping, remained the exception as linkage approaches dominated. Studies of many pairs of relatives (most commonly, affected sib pairs) were especially prevalent owing to the ease of collecting such samples versus samples from extended families, to their theoretical appeal for mapping complex traits[16–18] and to the availability of powerful analysis tools[19]. These genome scans were carried out for many common diseases that show complex inheritance, but they failed to find many reproducible loci. With these findings, the initial belief that a few major genes would explain susceptibility to complex diseases gave way to the realization that the level of complexity was much

higher and that many loci of individually small effect were involved. Because such loci are difficult to identify by family linkage owing to limited power, the search was on for new approaches.

Modern proposals

In an influential perspective, Risch and Merikangas argued that association studies should be more powerful than family linkage studies for detecting high-frequency, small-effect polymorphisms[20]. Linkage studies rely on allele sharing by descent among affected relatives, and their low power to detect such polymorphisms is due to two factors. First, when the increased risk conferred by an allele is small, some relatives will be affected because of other causes and will not carry the risk allele. Second, when an allele is common, it can enter the family through multiple founders, erasing clear inheritance patterns. These effects combine to decrease sharing by descent to the point at which an impractically large number of families must be studied to detect it. Association studies still suffer from the first effect (which is inherent to searches for small effects) but not the second, and therefore they have a higher sensitivity in detecting common variants with small effects. The increase in power is such that even testing large numbers of polymorphisms, with the ensuing statistical costs of multiple testing, does not erase the advantage of association studies[20,21]. In addition to offering higher power with the same sample sizes, association studies also have the practical advantage that large samples of unrelated cases and controls can be collected much more easily than family-based samples.

Risch and Merikangas issued a call for a catalogue of all variants in human genes, and set out a challenge "to the molecular technologists to develop the tools" for their identification and genotyping[20]. This call was echoed by Eric Lander, who hypothesized that common variants of modest effect might hold the key to susceptibility to common diseases[21] (this was subsequently codified as the common disease–common variant hypothesis). Lander also noted that the role of noncoding variation might be studied by the use of LD mapping with a sufficiently dense polymorphism map.

These proposals were formalized the following year in a policy forum by Collins, Guyer and Chakravarti[22]. They made explicit the distinction between the direct approach of cataloguing all common functional variants and the indirect approach of relying on a dense map of SNPs for LD mapping (FIG. 2). A back-of-the-envelope calculation put the likely size of shared ancestral haplotypes in the range of 10–100 kb, leading to a proposal to identify at least 100,000 SNPs. To achieve this goal, The SNP Consortium, a public–private partnership, was launched in 1999.

Charting the course

The early proposals for genome-wide studies were audacious, because the number of SNPs known at the time was small, and the approaches to their discovery and genotyping were cumbersome. In 1998, Wang and colleagues performed an important feasibility study, discovering some 3,000 SNPs and developing an array-based genotyping approach that could assay hundreds of SNPs in parallel[23]. The SNP consortium and the HapMap project would eventually bridge the gap between this early survey and the much larger number of SNPs required.

How many SNPs are needed? The number of SNPs that are required for LD mapping obviously depends on the genomic extent of LD because genotyped SNPs must be spaced sufficiently densely to be in LD with most of the (potentially disease-associated) variants that are not genotyped. At the time the proposals for GWAS were made, few empirical estimates of the extent of LD were available, and these varied wildly from observations of LD over hundreds of kb to the breakdown of LD at very short distances. This range of observations translated into an uncertainty of up to three orders of magnitude in the required number of SNPs — from thousands to millions. Even several years later, the number of SNPs required for GWAS was said to be in the range of 30,000–1,000,000, based on a survey of empirical studies[24]. Starting in 1997, I attempted to reduce this uncertainty by using simple population-genetics models to calculate the likely extent of LD. A highly realistic model could not be constructed at the time because of a lack of detailed information regarding both the demographic history of different populations and the variation in recombination rate at short distances. Instead, the aim was to obtain a reasonable estimate. In the model that was designed to approximate the global human population, moderate levels of LD were confined to regions of approximately 6 kb, thus leading to the prediction that some 500,000 SNPs would be required for GWAS, even if relatively low LD levels between mapped SNPs and functional variants were deemed acceptable[25]. The predicted number of SNPs was considerably larger than the goals of SNP discovery projects at the time[26], and led to an increase in the targeted number. Indeed, less than 2 years later, a map of 1 million SNPs was reported[24]. Given the simplified nature of the model that was used to calculate the estimate of 500,000 SNPs, this number has held up remarkably well — most of the recent successes of GWAS came when approximately this number of SNPs could be genotyped within individual studies, and the current generation of commercial SNP-typing products deploys some 500,000–1,000,000 SNPs.

When the prediction is viewed from the vantage point of the extensive empirical data available today (for example, REF 27),

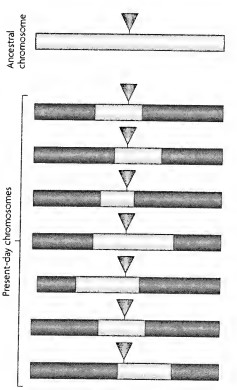


Figure 1 | **Linkage disequilibrium around an ancestral mutation.** The mutation is indicated by a red triangle. Chromosomal stretches that are derived from the common ancestor of all mutant chromosomes are shown in light blue, whereas new stretches introduced by recombination are shown in dark blue. Markers that are physically close (that is, within the light-blue regions of present-day chromosomes) tend to remain associated with the ancestral mutation, even as recombination whittles down the region of association over time.

Box 1 | **A brief history of genetic markers**

Human genetic mapping was initially based on restriction fragment length polymorphisms (RFLPs)[6,9,49,50] — fragment length variants generated through the presence and/or absence of restriction enzyme recognition sites[6,13,53,54]. RFLPs resulted from various sequence changes including base substitutions, insertions and deletions, and were laboriously assayed by Southern blots. Southern blots were superseded by PCR-based assays for microsatellite markers (also known as short tandem repeats or simple sequence length repeats)[10,51]. Microsatellites are di-, tri- or tetranucleotide repeat sequences that are composed of many tandem repeats. Many alleles are generally associated with each microsatellite within most populations, hence their use as markers for carrying out family-based linkage analysis[12,51]. More recently, SNPs have become the markers of choice; their lower polymorphism is offset by their abundance and ease of genotyping[13,56], and their low mutation rates make them especially suitable for linkage disequilibrium mapping.

it is clear that the actual average extent of LD is greater than in the model. This is especially true in non-African populations, most likely owing to a combination of demographic factors and a clustering of recombination events at hot spots[28] (both of these effects were anticipated when the prediction was made[25]). However, the calculation of the number of SNPs assumed both a relatively low acceptable level of LD and coverage of each region of LD with a single SNP. In practice, GWAS have set a higher standard for required LD, and multiple SNPs are used to 'tag' each region of high LD. These factors combined lead to the most current empirical estimates of approximately 500,000 SNPs for non-African and 1,000,000 SNPs for African populations to ensure adequate coverage of the genome in GWAS, even when high LD levels between mapped SNPs and functional variants are required[29].

How should the SNPs be chosen? Initially, the discussion focused on simply assembling a dense collection of SNPs. However, both theoretical considerations and early empirical studies suggested that the physical extent and the local patterns of LD were likely to vary across the genome and among populations. In commenting on one empirical study[30] in 1999, I proposed that LD among a dense collection of SNPs be measured empirically across the genome and in different populations in order to identify the most efficient SNP panels for association studies (FIG. 3); such panels would vary in their density by region of the genome and by population[31]. The result of such empirical studies would constitute an LD map of the human genome[31]. As SNP discovery efforts continued, empirical data confirmed both the need for hundreds of thousands of SNPs and the fact that these SNPs could not be chosen at random or by uniform spacing across the genome[5,32–35]. Rather, a million or more SNPs would need to be genotyped in substantial numbers of individuals from multiple populations in order to select sets of several hundred thousand (with the precise number depending on the population) that would efficiently capture untyped common variants[5,32–34]. These observations gave rise to the HapMap project and a parallel effort by Perlegen Sciences, which eventually joined forces to produce the SNP panels that are being used today[29,36,37]. These projects also drove the development of rapid and cost-effective genotyping technologies, setting the stage for GWAS. These recent developments are well chronicled elsewhere (for example, REF. 38).

The road ahead

In the past two decades, GWAS have progressed from visionary proposals, made when neither the sequence of the human genome nor many variations in this sequence were known, to routine practice of screening 500,000–1,000,000 SNPs in thousands of individuals. The recently reported phase 2 of the HapMap[29] now includes 3 million SNPs, estimated to cover one-quarter to one-third of all human SNPs with frequencies above 5%. Where do we go from here?

The recent crop of discoveries from GWAS is a major advance in our understanding of the genetic basis of common diseases, as well as normal human variation[39,40]. Nevertheless, the associated loci that have been identified usually have small individual effects on phenotype, and even collectively tend to explain only a small fraction of the heritable component[3]. For some diseases studied, no significant loci have been identified[3,41]. This failure to detect loci that explain the bulk of the heritable components of the phenotypes studied could be attributable to several factors. First, because the detected loci have small effects, the power to detect them is low, and more such loci remain to be discovered as sample sizes increase. Second, association studies can only detect the effects that are due to relatively common alleles. Rare alleles remain to be discovered — both at the loci that are identified by GWAS because they also have common alleles with phenotypic effects, and at other loci that do not have such common alleles. The former can be found by focused resequencing studies of the loci identified by GWAS; finding the latter might require resequencing of other genes in the relevant pathways, of the exons of all genes[42–44] or of the entire genome. Third, we might be missing the effects of structural variation, of other less well-studied types of genome alterations[45], and of interactions among variants and between genetic and environmental factors.

It is only a matter of time before all SNPs with appreciable frequencies in the human population have been discovered.

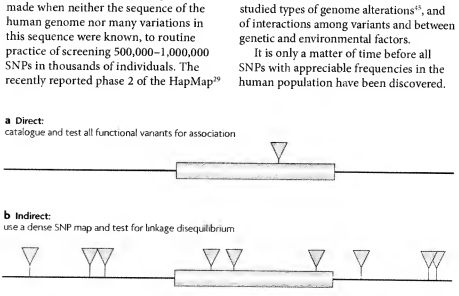


Figure 2 | **Alternative designs for genome-wide association studies. a** | Direct approach of testing a catalogue of all common functional variants in the genome. **b** | Indirect approach of testing a dense map of SNPs and relying on linkage disequilibrium to detect associations that are due to untested functional variants.

Indeed, efforts to discover and genotype additional SNPs in larger and more diverse population samples are underway. Assuming that the genotyping technologies can keep up the pace, the indirect association studies relying on LD will be replaced by direct association studies that assay all relatively common SNPs (perhaps the estimated 11 million SNPs with minor allele frequencies exceeding 1% in the population[66]), although it will probably still be worthwhile to exclude wholly redundant SNPs. Thus, LD and haplotype maps are merely useful but temporary shortcuts. An interesting finding of phase 2 of the HapMap is that for approximately 1% of all SNPs (tens of thousands), the basic assumption of indirect association mapping breaks down — these SNPs are not in LD with any others, often owing to their location in hot spots of recombination, and are thus 'untaggable' and must be assayed directly for phenotypic associations[39]. Tailored approaches that are under development will cover structural variants[67]. Studies based on the resequencing of individual genomes (rather than genotyping of known variants) will be needed to begin to comprehensively address the role of rare variants and *de novo* mutations, and will eventually replace genotyping studies altogether, although this is likely to take some time. It is worth noting that resequencing studies of rare variants have to rely on the recognition of many different variants, each of which alters the function of the same gene or pathway in different individuals. Whereas recognition of likely functional variants in coding regions is straightforward, detecting functional changes in noncoding DNA poses a major challenge. This is because regulatory sequences can be located far from the coding region and are often difficult to identify, and we do not have a ready connection between nucleotide differences and function for these sequences.

Looking further ahead, we can already envision the day when the genome sequences of a significant fraction of the population are known, at least in the developed world. Assuming that the relevant logistical and ethical issues can be solved, what will we learn from combining this unprecedented scope of genetic information with medical records and other phenotypic data? We are just beginning to get the first glimpses of the real underlying genetic complexity of phenotypic variation. Complexity is present on multiple levels, and might be fruitfully thought of as 'fractal'. First, many loci are involved; we do not yet know how many but the number could be in the hundreds for many traits. Second, individual loci can often represent variation in multiple linked genes, as has been found in model organisms (for example, REF. 48). Third, each gene is likely to contain multiple functional variants, including both 'superalleles' of linked alterations on one haplotype and allelic series with a range of allele frequencies and effect sizes. Non-additive interactions can be present at all levels. GWAS detect effects at the locus level, and an important challenge for future studies is identification of the genes, the functional variants and the functional mechanisms underlying phenotypic associations. Currently, such studies require painstaking, low-throughput experiments in cell lines and animal models.

It is possible that some genetic contributions to human phenotypic variation might be too subtle to unravel, even when our surveys of the genome become truly comprehensive and the sample sizes approach that of the human population. Aside from the question of how much of the population variation we will ultimately be able to explain, we also have to ask how we can piece together individual risk from many small genetic contributions. Will we ultimately be able to classify individuals into meaningful groups with regard to risk of specific common diseases or response to drugs, as envisioned in personalized medicine? Doubtless this will be (or already is) true in some cases, but it is currently unknown how general such classifications are. We might need to replace some current phenotypic and disease classifications with ones that better correspond to the underlying genetic causes, perhaps by developing methods to iteratively refine phenotypic categories by combining genotypic and phenotypic information. Careful and detailed measures of phenotypes and environmental exposures will also have an important role. Clearly, we have a lot of work to do before an individual genome sequence is more phenotypically informative than it is today[49,50]. In the meantime, great care is required in offering genome-based information to individuals[51,52].

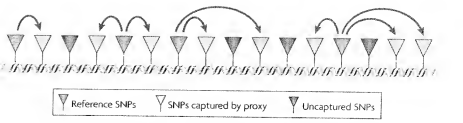


Figure 3 | **Schematic of a genomic region to be tested for association with a phenotype.** The four reference SNPs in the mapping panel are indicated by red triangles: these are genotyped directly. The eight SNPs indicated by yellow triangles are captured through linkage disequilibrium (by proxy) with the reference SNPs denoted by arrows. The four SNPs indicated by blue triangles are neither genotyped nor in linkage disequilibrium with the reference SNPs; phenotypic association that is due to one of these would be missed.

Concluding remarks

What is the best future direction for human genetics? There are essentially three avenues to pursue: much larger samples; better assays of genome variation that can capture both common alterations that are not in LD with SNP panels and rare variants; and more detailed phenotyping. Undoubtedly, each of these approaches has a role, and we do not yet have all the information needed to decide which will prove most fruitful. Therefore, it is a high priority to apply a full battery of approaches to several model diseases and phenotypes in order to empirically determine the range of outcomes, just as the Wellcome Trust Case Control Consortium study of seven diseases provided an empirical guide for GWAS[41]. In my opinion, the most pressing question is the contribution of rare variants, both in the genes that harbour common risk variants and in those that do not. This question is also the most difficult to address comprehensively with today's technologies, but it seems imperative that we prioritize studies to begin to get the answers.

Leonid Kruglyak is at the Lewis-Sigler Institute for Integrative Genomics and the Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey 08544, USA.

e-mail: leonid@genomics.princeton.edu

doi:10.1038/nrg2316

Published online 19 February 2008

1. Altshuler, D. & Daly, M. Guilt beyond a reasonable doubt. *Nature Genet.* **39**, 813–815 (2007).
2. Bowcock, A. M. Genomics: guilt by association. *Nature* **447**, 645–646 (2007).
3. Gibson, G. & Goldstein, D. B. Human genetics: the hidden text of genome-wide associations. *Curr. Biol.* **17**, R929–R932 (2007).
4. Topol, E. J., Murray, S. S. & Frazer, K. A. The genomics gold rush. *JAMA* **298**, 218–221 (2007).
5. The International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
6. Botstein, D., White, D. L., Skolnick, M. & Davis, R. W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* **32**, 314–331 (1980).
7. Solomon, E. & Bodmer, W. F. Evolution of sickle variant gene. *Lancet* **1**, 923 (1979).
8. Lander, E. S. & Botstein, D. Mapping complex genetic traits in humans: new methods using a complete RFLP linkage map. *Cold Spring Harb. Symp. Quant. Biol.* **51**, 49–62 (1986).
9. Kan, Y. W. & Dozy, A. M. Polymorphism of DNA sequence adjacent to human beta-globin structural gene: relationship to sickle mutation. *Proc. Natl Acad. Sci. USA* **75**, 5631–5635 (1978).
10. Bodmer, W. F. Human genetics: the molecular challenge. *Cold Spring Harb. Symp. Quant. Biol.* **51**, 1–13 (1986).
11. Donis-Keller, H. *et al.* A genetic linkage map of the human genome. *Cell* **51**, 319–337 (1987).
12. Weissenbach, J. *et al.* A second-generation linkage map of the human genome. *Nature* **359**, 794–801 (1992).
13. Kerem, B.-S. *et al.* Identification of the cystic fibrosis gene: genetic analysis. *Science* **245**, 1073–1080 (1989).
14. Houwen, R. H. J. *et al.* Genome scanning by searching for shared segments: mapping a gene for benign recurrent intrahepatic cholestasis. *Nature Genet.* **8**, 380–386 (1994).
15. Puffenberger, E. G. *et al.* Identity-by-descent and association mapping of a recessive gene for Hirschsprung disease on human chromosome 13q22. *Hum. Mol. Genet.* **3**, 1217–1225 (1994).
16. Risch, N. Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs. *Am. J. Hum. Genet.* **46**, 242–253 (1990).
17. Risch, N. Linkage strategies for genetically complex traits. II. The power of affected relative pairs. *Am. J. Hum. Genet.* **46**, 229–241 (1990).
18. Risch, N. Linkage strategies for genetically complex traits. I. Multilocus models. *Am. J. Hum. Genet.* **46**, 222–228 (1990).
19. Kruglyak, L. & Lander, E. S. Complete multipoint sib-pair analysis of qualitative and quantitative traits. *Am. J. Hum. Genet.* **57**, 439–454 (1995).
20. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
21. Lander, E. S. The new genomics: global views of biology. *Science* **274**, 536–539 (1996).
22. Collins, F. S., Guyer, M. S. & Chakravarti, A. Variations on a theme: cataloging human DNA sequence variation. *Science* **278**, 1580–1581 (1997).
23. Wang, D. G. *et al.* Large-scale identification, mapping and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**, 1077–1082 (1998).
24. The International SNP Map Working Group. A map of human genome sequence variation containing 1 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
25. Kruglyak, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genet.* **22**, 139–144 (1999).
26. Collins, F. S. *et al.* New goals for the US human genome project: 1998–2003. *Science* **282**, 682–689 (1998).
27. Pe'er, I. *et al.* Biases and reconciliation in estimates of linkage disequilibrium in the human genome. *Am. J. Hum. Genet.* **78**, 588–603 (2006).
28. Reich, D. E. *et al.* Human genome sequence variation and the influence of gene history, mutation and recombination. *Nature Genet.* **32**, 135–142 (2002).
29. The International HapMap Consortium. A second generation human haplotype map of over 3 million SNPs. *Nature* **449**, 851–861 (2007).
30. Lonjou, C., Collins, A. & Morton, N. E. Allelic association between marker loci. *Proc. Natl Acad. Sci. USA* **96**, 1621–1626 (1999).
31. Kruglyak, L. Genetic isolates: separate but equal? *Proc. Natl Acad. Sci. USA* **96**, 1170–1172 (1999).
32. Carlson, C. S. *et al.* Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. *Nature Genet.* **33**, 518–521 (2003).
33. Gabriel, S. B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229 (2002).
34. Reich, D. E., Gabriel, S. B. & Altshuler, D. Quality and completeness of SNP databases. *Nature Genet.* **33**, 457–458 (2003).
35. Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J. & Lander, E. S. High-resolution haplotype structure in the human genome. *Nature Genet.* **29**, 229–232 (2001).
36. Hinds, D. A. *et al.* Whole-genome patterns of common DNA variation in three human populations. *Science* **307**, 1072–1079 (2005).
37. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
38. Hirschhorn, J. N. & Daly, M. J. Genome-wide association studies for common diseases and complex traits. *Nature Rev. Genet.* **6**, 95–108 (2005).
39. Sulem, P. *et al.* Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nature Genet.* (2007).
40. Weedon, M. N. *et al.* A common variant of *HMGA2* is associated with adult and childhood height in the general population. *Nature Genet.* **39**, 1245–1250 (2007).
41. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
42. Albert, T. J. *et al.* Direct selection of human genomic loci by microarray hybridization. *Nature Methods* **4**, 903–905 (2007).
43. Hodges, E. *et al.* Genome-wide *in situ* exon capture for selective resequencing. *Nature Genet.* (2007).
44. Porreca, G. J. *et al.* Multiplex amplification of large sets of human exons. *Nature Methods* **4**, 931–936 (2007).
45. Legendre, M., Pochet, N., Pak, T. & Verstrepen, K. J. Sequence-based estimation of minisatellite and microsatellite repeat variability. *Genome Res.* **17**, 1787–1796 (2007).
46. Kruglyak, L. & Nickerson, D. A. Variation is the spice of life. *Nature Genet.* **27**, 234–236 (2001).
47. Estivill, X. & Armengol, L. Copy number variants and common disorders: filling the gaps and exploring complexity in genome-wide association studies. *PLoS Genet.* **3**, 1787–1799 (2007).
48. Sinha, H., Nicholson, B. P., Steinmetz, L. M. & McCusker, J. H. Complex genetic interactions in a quantitative trait locus. *PLoS Genet.* **2**, e15 (2006).
49. Brenner, S. E. Common sense for our genomes. *Nature* **449**, 783–784 (2007).
50. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254 (2007).
51. Anonymous. Risky business. *Nature Genet.* **39**, 1415 (2007).
52. McGuire, A. L., Cho, M. K., McGuire, S. E. & Caulfield, T. Medicine. The future of personal genomics. *Science* **317**, 1687 (2007).
53. Gusella, J. F. *et al.* A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* **306**, 234–238 (1983).
54. Wyman, A. R. & White, R. W. A highly polymorphic locus in human DNA. *Proc. Natl. Acad. Sci. USA* **77**, 6754–6758 (1980).
55. Weber, J. L. & May, P. E. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am. J. Hum. Genet.* **44**, 388–396 (1989).
56. Kruglyak, L. The use of a genetic map of biallelic markers in linkage studies. *Nature Genet.* **17**, 21–24 (1997).

Acknowledgements

I thank many colleagues over the years, the participants at the 2007 Banbury Center Meeting "From Statistics To Genes: Figuring Out the Molecular Basis of Complex Traits", D. Altshuler for discussions, and D. Botstein, A. Chakravarti, B. Coller, D. Goldstein and L. Rosenberg for comments on the manuscript. I regret that space constraints prevented me from citing other important work in the field. Supported by a MERIT award from the National Institutes of Health (R37 MH059520) and a James S. McDonnell Centennial Fellowship in Human Genetics.

DATABASES

OMIM: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM
benign recurrent intrahepatic cholestasis (BRIC) | Hirschsprung disease

FURTHER INFORMATION

Leonid Kruglyak's homepage: http://www.genlab.princeton.edu/index.php?option=content&task=view&id=217
Perlegen Sciences: http://www.perlegen.com
The International HapMap project: http://www.hapmap.org

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

Kingsmore, S.F., et al., Nature Reviews Genet 7:221-230 (2008)

Genome-wide association studies: progress and potential for drug discovery and development

Stephen F. Kingsmore, Ingrid E. Lindquist, Joann Mudge, Damian D. Gessler and William D. Beavis

Abstract | Although genetic studies have been critically important for the identification of therapeutic targets in Mendelian disorders, genetic approaches aiming to identify targets for common, complex diseases have traditionally had much more limited success. However, during the past year, a novel genetic approach — genome-wide association (GWA) — has demonstrated its potential to identify common genetic variants associated with complex diseases such as diabetes, inflammatory bowel disease and cancer. Here, we highlight some of these recent successes, and discuss the potential for GWA studies to identify novel therapeutic targets and genetic biomarkers that will be useful for drug discovery, patient selection and stratification in common diseases.

Genetic linkage
Co-segregation (reduced recombination) of a trait and an allele in related subjects (pedigrees) more often than explicable by chance.

Dominant
An allele that confers a trait even when it is heterozygous (present as a single copy in a genome).

Recessive
An allele that confers a trait only when it is homozygous (present in two copies in a genome, one from each parent).

Endophenotype
A measurable component of a phenotype.

National Center for Genome Resources, 2935 Rodeo Park Drive East, Santa Fe, New Mexico 87505, USA. Correspondence to S.F.K. e-mail: sfk@ncgr.org
doi:10.1038/nrd2516
Published online
15 February 2008

Genetic factors are known to have an important role in many common diseases, and the identification of genetic determinants for such diseases has the potential to provide insights into disease pathogenesis, revealing novel therapeutic targets or strategies. Genetic factors could also provide useful biomarkers for diagnosis, patient stratification and prognostic or therapeutic categorization. In addition, given that inherited genetic factors are present at birth, knowledge of these factors could facilitate timely preventative or ameliorative interventions.

During the past 25 years, genetic linkage-based studies have proved very effective in identifying causal genetic factors in Mendelian (single gene) disorders; causal genes for more than 1,300 dominant and recessive Mendelian diseases have been identified[1]. Most common diseases and endophenotypes, however, do not exhibit Mendelian inheritance, but rather feature complex, multifactorial expression and inheritance. Although linkage-based methods have been broadly applied, these studies have had little success in identifying the allelic determinants of common disorders[2]. In particular, there has been poor replication among studies, whereby an initial study identifies an allele (genotype) with large estimated genetic effects (relative risk) but subsequent studies fail to corroborate the results[3,4]. In part, this reflects the dependence of linkage-based studies on unusually informative families (with multiple affected and unaffected individuals), which induce a bias toward rare, semi-Mendelian disease subsets in subpopulations. Reports of successful identification of genetic variants in common diseases using an approach that circumvents this limitation — genome wide association (GWA) studies — have therefore generated considerable excitement.

Human GWA studies are based on three hypotheses: First, the common trait/common variant hypothesis proposes that the genetic architecture of complex traits consists of a limited number of common alleles, each conferring a small increase in risk to the individual[5,6]; second, the brief history of most human populations precludes sufficient generations (or meioses) to create recombination events (or mutations) between closely located, common (ancient) variants; and, third, suppression of meiotic recombination (coldspots) occurs very frequently. Thus, approximately 80% of the human genome is comprised of around 10 kb regions that exhibit reduced recombination in human populations (haplotypes)[7]. Genetic variants (alleles) within haplotypes are in linkage disequilibrium (LD). This phenomenon enables much of the recombination history in a population to be ascertained by genotyping a large set of well-spaced, common (ancient) variants throughout the genome, especially if variant selection is informed by knowledge of haplotypes. During the last 10 years, more than 10 million single nucleotide polymorphisms (SNPs) have been identified[8]. Furthermore, the International HapMap project has genotyped approximately 4 million

Box 1 | Useful resources and databases for genetic-based studies

- **Genetic Association Database:** An archive of human genetic association studies of complex diseases. http://geneticassociationdb.nih.gov/
- **Schizophrenia Gene Database:** An archive of genetic association studies performed on schizophrenia phenotypes. http://www.schizophreniaforum.org/res/sczgene/default.asp
- **Online Mendelian Inheritance in Man:** A catalogue of human genes and genetic disorders. http://www.ncbi.nlm.nih.gov/sites/entrez?db=OMIM&itool=toolbar
- **Human Gene Mutation Database.** A catalogue of published gene lesions responsible for human inherited disease. http://www.hgmd.cf.ac.uk/ac/index.php
- **Human Genome Variation Database:** A catalogue of normal human gene and genome variation. http://www.hgvbase.org/
- **dbSNP:** A catalogue of human single nucleotide polymorphisms. http://www.ncbi.nlm.nih.gov/projects/SNP/
- **GeneSNPs:** A database of polymorphisms in human genes that are thought to have a role in susceptibility to environmental exposure. http://www.genome.utah.edu/genesnps/
- **PharmGKB:** A database of pharmacogenomics research. http://www.pharmgkb.org/index.jsp
- **GeneCards:** A database of human genes that includes genomic, proteomic and transcriptomic information, as well as orthologies, disease relationships, SNPs, gene expression and gene function. http://www.genecards.org/

Multifactorial
Inheritance of a trait that is attributable to two or more genes and their interaction with the environment (also known as polygenic inheritance).

Allele
The DNA code at a given locus (position) on a chromosome.

Genome-wide association study
A comprehensive search of the human genome for genetic risk factors for a trait by a case–control association study involving comparisons of hundreds of thousands of alleles between unrelated subjects with and without a trait.

Haplotype
A combination of alleles at linked loci (on a single chromatid) that are transmitted together more often than explicable by chance.

Linkage disequilibrium
(LD). Combinations of alleles in a population that differ in frequency from that expected from random formation of haplotypes from alleles based on their frequencies.

Minor-allele frequency
The allele frequency of the less frequently occurring allele of a polymorphism.

Case–control association study
Comparison of the frequency of an allele between unrelated subjects with and without a trait. A difference in allele frequency between the two groups indicates that the allele might change the likelihood of the trait.

Genetic association
Correlation of a trait and an allele in a population more often than explicable by chance.

Genocopy
A genotype at a locus that produces a phenotype that is indistinguishable from that produced by a genotype at another locus.

Phenocopy
An environmentally produced phenotype that simulates the effect of a particular genotype

common SNPs (occurring with a minor-allele frequency of more than 5%) in human populations and has assembled these genotypes computationally into a genome-wide map of SNP-tagged haplotypes[7]. These resources, together with array technologies for massively parallel SNP genotyping and the well-established epidemiological case-control association studies have rendered GWA feasible (BOX1,FIG.1).

Initial genetic association studies focused on candidate loci and exhibited a lack of replication among studies[8,10]. There were biological explanations for inconsistent results: unobserved, confounding biological sources of heterogeneity, including inconsistent or poorly defined measurements of the phenotype, heterogeneous genetic sources for the phenotype (genocopies), population stratification (ethnic ancestry), population-specific LD, heterogeneous genetic and epigenetic backgrounds or heterogeneous environmental influences (phenocopies). In addition, there were statistical reasons for irreproducibility, including failure to control the rate of false discoveries, model misspecification and heterogeneous bias in estimated effects among studies[11–14]. Also, a frequent source of non-replication was lack of power due to the limited number of individuals genotyped and phenotyped[15,16].

In order to ameliorate poor replication, GWA experiments employ multi-tiered experimental designs with discovery, replication and biological validation stages[17] (FIG. 1). Tiered designs are critical for cost-effective detection of meaningful, hypothesis-generating, genotype–phenotype associations given the large number of comparisons involved, prior probability estimates of association, sample sizes, resampling procedures and statistical significance thresholds. GWA studies also owe their statistical power to their large cohort size and high rate of SNP detection. Currently, a respected threshold for uncorrected, significant associations is $P < 5 \times 10^{-7}$ (REFS 18,19). Alleles with moderately less significant associations, however, are often also reported, as they might indicate loci that reach the aforementioned threshold in subsequent studies.

Results of initial GWA studies

The first GWA study, published in 2002, evaluated acute myocardial infarction (AMI)[20]. The discovery, or nomination, phase comprised the examination of genotype–phenotype association signals in 65,671 coding domain SNPs (cSNPs) in 752 cases and controls (TABLE 1). Although subsequent studies have used up to 20 times this number of non-coding SNPs, gene-tagging SNPs are more informative, as the majority of true-positive associations are expected to be with genes[1]. Even more informative are screens that employ functional cSNPs, such as nonsynonymous SNPs (nsSNPs), that are candidate, causal (risk-enhancing) gene alleles[1,21–28]. The replication, or confirmatory, phase examined associations of 26 SNPs in 2,137 individuals and confirmed association of AMI with a 50 kb region containing lymphotoxin-α (*LTA*), nuclear factor of kappa light polypeptide gene enhancer in B cells (also known as *RELA*), nuclear factor of kappa light polypeptide gene enhancer in B cells inhibitor-like 1 (*NFKBIL1*) and human leukocyte antigen (HLA)-B associated transcript 1 (*BAT1*) genes. Additional replication studies have been undertaken, some of which have confirmed an association of this region with AMI-related phenotypes and, in particular, one nsSNP in *LTA*[29–35]. The association of *LTA* with AMI was an unexpected finding, suggesting a novel therapeutic target.

A second, pioneering GWA study examined age-related macular degeneration (AMD)[36] (TABLE 1). The discovery phase sought associations of 105,980 SNPs with AMD in 96 cases and 50 control individuals. Despite the small cohort size, SNPs in the complement factor H (*CFH*) gene, including an nsSNP, showed significant association with AMD. Replication was not performed, but subsequent studies have replicated associations of *CFH* alleles with AMD[37–49]. Of all common diseases examined by GWA to date, AMD is unique in that a single haplotype explains 61% of the genetic variance, conferring a homozygous odds ratio of 7.4. To put this in perspective, this is of a similar magnitude to the classic associations of *HLA-B27* with anterior uveitis/ankylosing spondylitis and *HLA* alleles with type 1 diabetes mellitus (T1DM). Complement

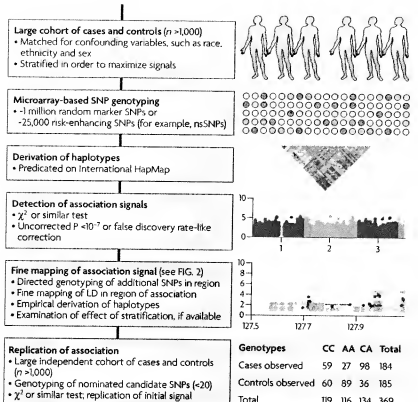


Figure 1 | Overview of the general design and workflow of a genome-wide association (GWA) study. The discovery phase entails genotyping many case and control DNA samples and evaluation for significant associations. The replication phase involves fine mapping of association signals and independent confirmation in a second cohort. Biological validation is important for translation of GWA findings into diagnostic or therapeutic discoveries.

pathway dysregulation was a novel, unexpected association with AMD. Subsequent studies have shown an association of AMD with two additional members of the alternative complement pathway (factor B (*CFB*) and *C3*)[41,42]. These findings, together with biological validation studies, have led to the initial development of new AMD therapies, based upon complement inhibition.

In the past year, technical challenges associated with GWA were largely overcome, genotyping costs were decreased and a significant number of studies have used SNP genotyping arrays in larger population groups to produce replicated associations between individual SNP alleles and common diseases.

Inflammatory bowel disease. Five large GWA studies have examined Crohn's disease and ulcerative colitis, two histologically distinct types of inflammatory bowel disease (IBD) (TABLE 1). Four of the studies used microarrays featuring between 300,000 and 400,000 SNPs[18,43–45], whereas the fifth study genotyped approximately 16,000 nsSNPs[24]. Two follow-up studies sought to replicate the most significant signals from the Wellcome Trust Case Control Consortium (WTCCC) study[18], one in a European population and another in a Japanese population[46,47]. The European study replicated significant signals of the WTCCC study, but some of the alleles failed to reach significance in the Japanese study and others were not detected. The failure to replicate signals in different studies might reflect true differences between populations, differences in phenotype ascertainment or a lack of power.

Considering the six studies of European populations, there was significant replication of specific allele associations with Crohn's disease (TABLES 1,2). Three associations were concordant in four out of five studies (representing the genes caspase recruitment domain 15 protein (*CARD15*, also known as *NOD2*), interleukin 23 receptor (*IL23R*) and ATG16 autophagy related 16-like 1 (*ATG16L1*)). Of note, *CARD15* had previously been identified as a susceptibility gene by linkage-based approaches[48,49]. One gene, prostaglandin E receptor 4 (*PTGER4*) showed association in two out of five studies. In addition, several disease-associated intergenic segments have been replicated. IBD susceptibility genes that have been identified to date appear to coalesce into biological networks involving innate immunity, autophagy and phagocytosis[50]. In addition, alleles of two genes associated with Crohn's disease (*IL23R* and *PTPN2*) have shown association with other autoimmune disorders[51,52], suggesting the existence of autoimmune susceptibility 'supergenes'. There is great interest in alleles that exhibit pleiotropic associations, as they potentially represent blockbuster targets that cross-over therapeutic categories (TABLE 2).

In common with most GWA studies to date, estimated genetic effects (relative risks) of IBD-associated loci are small[18,46]. However, as many of these variants were common, the population attributable risk — an estimate of the percentage of cases of disease that would be avoided if the allele(s) were absent — was substantial. Of several studies that looked for epistatic interactions between IBD association signals, two found suggestive evidence of epistasis involving two different pairs of genes[24,52].

Diabetes mellitus. A good example of the capabilities and limitations of GWA studies is type 2 diabetes mellitus[18,53–57] (T2DM; TABLE 1). Two studies examined association both with SNPs and haplotypes in the discovery phase[54,57]. Haplotype-based analysis can be more powerful than marker-by-marker analysis in association studies[22,58–61]. For example, haplotypes can correlate a specific phenotype with a specific gene in a small population sample even when individual SNPs cannot[62]. Case-control and family-based association studies were employed in several studies of T2DM.

The replication phases of these studies were impressive; two of them included over 9,000 replication individuals[53,54]. One study sought to replicate signals identified by the WTCCC study[18] by genotyping the most significant SNPs; 9 of 77 candidate SNPs reached a P $<5\times10^{-7}$ significance level[63]. The eight genes represented by these

Non-synonymous SNP (nsSNP)
A SNP that leads to a change in the amino acid sequence of the gene's resulting protein and that might therefore affect its function.

Odds ratio
A measure of risk that compares the probability of occurrence of a disease in a group with a risk allele with the probability in a control group.

Pleiotropy
A single gene that influences multiple phenotypic traits.

Epistasis
Modification of the action of a gene by another gene

SNPs were replicated in at least one other independent study (TABLES 1–3). The concordance of T2DM-associated genes between GWA studies is striking: of 10 novel associations, only two were unique to a single study.

Reassuringly, some of the genes identified by GWA in studies of TD2M have previously been associated with the disease in other types of genetic studies. For example, transcription factor 7-like 2 (*TCF7L2*) had previously

Table 1 | **Discovery and replication designs of recent GWA studies**

Disease	Discovery Phase			Replication Phase			Refs
	Number of individuals examined	Number of SNPs	Population	Number of individuals examined	Number of SNPs validated/ tested	Population	
AMD	146	105,980	Caucasian	96	2/50	Same	36
Asthma	2,642	307,328	UK/German	2,320	0/9	German	92
Atrial fibrillation	5,026	316,515	Icelandic	17,810	2/18	Icelandic/ European	94
Bipolar disorder	1,024 (pooled)	555,235	Western European	1,648	1/37	Same	91
Breast cancer	754	227,876	European	45,426	7/30	Same	85
	2,287	528,173	European	3,848	1/8	Same	19
	13,163	311,524	Icelandic	7,968	2/9	Various	86
Celiac disease	2,200	310,605	UK	2,480	5/27	Dutch/Irish	97
Colorectal cancer	1,890	547,647	Caucasian	23,121	2/18	Same	111
	2,593	99,632	Canadian	23,325	2/1,143	Same	112
Crohn's disease	1,923	304,413	European	2,150	4/37	Same	45
	1,103	16,360 nsSNP	German	2,670	3/72	Same	24
	1,475	302,451	Belgian	2,236	7/10	Same	44
IBD	1,095 + 834	308,332	European	2,885	10/27	Same	43
LOAD	1,086	502,627	Caucasian	ND	ND	ND	90
Lung cancer	673	116,204	Italian	621	0/1	Caucasian/ Norwegian	98
Memory	341 (pooled)	502,627	Swiss	680	½	Several	87
AMI	752	65,671 cSNPs	Japanese	2,137	4/26	Same	20
Nicotine dependence	548 (pooled)	2,427,357	European	1,929	0/31,960	European	93
Obesity	4,862	490,032	British/Irish	29,596	1/1	Same	71
Prolonged QT interval	3,966	88,500	German	4,451	1/7	European	96
Prostate cancer	4,517	316,515 & 243,957 haplotypes	Icelandic	3,655	2/2	Several	53
	12,791	310,520	Icelandic	5,050	2/5	European	78
	2,339	550,000	European	6,266	2/2	Several	79
RLS	2,045	236,758	European	2,336	9/13	Same	101
	15,970	306,937	Icelandic	2,206	1/70	Icelandic/US	100
T2DM	2,335	315,635	Finnish	2,473	10/80	Same	55
	4,900	393,453	European	9,103	10/77	Same	63
	7,805	313,179 & 339,846 haplotypes	Icelandic/Danish	3,382	2/47	Same	57
	1,316	392,935	French	5,511	8/57	Same	56
T2DM and Triglyceride levels	2,931	386,731 & 284,968 haplotypes	Finnish/Swedish	10,850	3/107	Several	54
T1DM	3,388	6,500 nsSNPs	European	12,229	1/1	Same	28
Bipolar disorder , Crohn's disease, T2DM, T1DM, HT, RA, CAD*	4,868	392,575	UK	ND	ND	ND	18

AMI, acute myocardial infarction; AMD, age-related macular degeneration; CAD, coronary artery disease; HT, hypertension; IBD, inflammatory bowel disease; LOAD, late-onset Alzheimer's disease; ND, not determined; nsSNP, non-synonymous SNP; RLS, restless leg syndrome; RA, rheumatoid arthritis; SNP, single nucleotide polymorphism; T1DM, type 1 diabetes mellitus; T2DM, type 2 diabetes mellitus. *There is overlap of the individuals genotyped in this study with REFS 58,63,80.

Table 2 | **Loci and variants associated with multiple diseases in GWA studies**

Locus	Variant (rs)	Disease	Refs
PTPN22	6679677	RA	18
		T1DM	18
IL2RA	2104286	RA	18
		T1DM	18
PTPN2	2542151	T1DM	51
		CD	18, 46
TCF2	4430796	T2DM	53
		PC	53
FTO	9939609	T2DM	18
		Obesity	71
APOE	4420638	Triglyceride level	54
		Alzheimer's disease	90
8q24	6983267	PC	79, 81, 113
		CC	111–113
IL23R	11209026	CD	43
		Psoriasis	21

Above variants are associated $P < 5 \times 10^{-7}$. CC, colorectal cancer; CD, Crohn's disease; PC, prostate cancer; RA, rheumatoid arthritis; T1DM, type 1 diabetes mellitus; T2DM, type 2 diabetes mellitus.

shown linkage to T2DM in the Icelandic population, and significant association in a candidate gene association study[64]. Heterozygous and homozygous carriers of *TCF7L2* risk alleles had relative risks of 1.45 and 2.41, respectively. TCF7L2 is a transcription factor that regulates the pro-glucagon gene in entero-endocrine cells[65]. *TCF7L2* alleles have also shown associations with endophenotypes such as a lower likelihood of response to the oral hypoglycaemic drug sulphonylurea[66] and increased risk of progression to T2DM among persons with impaired glucose tolerance[67].

In common with IBD, T2DM associations exhibited small estimated-effect sizes. Some of the candidate genes from GWA studies were consistent with biological processes that have previously been implicated in the pathogenesis of T2DM, such as pancreatic islet beta-cell function and insulin biosynthesis. However, these studies also suggested new components of these processes, such as zinc transport and Wnt-signalling[56,64,68]. Validation of T2DM candidate genes as therapeutic targets will require additional studies to identify causal susceptibility alleles and to determine their precise effect on cell biology.

Three studies performed initial modelling of how loci combine to affect susceptibility to T2DM[56,63,68]. One study found evidence of epistatic interactions between two genes. Otherwise, T2DM appeared to fit a polygenic threshold model with additive/multiplicative effects of individual loci. However, until the causal alleles that underpin these association signals have been found, it is not possible to make categorical statements about the allelic architecture of T2DM.

Frequencies of T2DM associated alleles showed considerable variation between ethnic and racial groups. Despite these differences, however, T2DM-associated risk alleles were conserved between independent populations, implying an ancient origin of these polymorphisms[70].

Expansion of an initial association of an allele with a categorical trait (such as the presence of a disease) with quantitative component phenotypes (endophenotypes) is an approach pioneered with apolipoprotein E (*APOE*) alleles in Alzheimer's disease. It appears to be highly instructive in elucidating the mechanism of action of alleles in disease pathogenesis. One T2DM GWA extended its analysis to a quantitative endophenotype: T2DM-related obesity (measured by body-mass index (BMI); TABLE 1)[54,71]. Alleles associated with T2DM in the fat-mass and obesity-associated gene (*FTO*)[18,55,63] also showed an association with BMI (TABLES 2,5). Association of *FTO* with obesity has since been confirmed[72].

Two GWA studies examined T1DM. One examined 6,500 nsSNPs[28] and the other evaluated 392,575 SNPs[18]. Four T1DM, susceptibility loci had previously been identified by linkage-based methods (class II MHC alleles, *CTLA4*, *PTPN22* and insulin). GWA studies replicated the association with *PTPN22* and identified several novel loci, including *C12orf30*, *KIAA0350* (also known as *CLEC16A*) and *IFIH1* (each replicated in two studies). Twenty-one T1DM candidate genes that have previously shown linkage or association are currently undergoing replication studies[73].

T1DM, like rheumatoid arthritis and IBD, is an autoimmune disorder. Medical practitioners have long noted familial aggregation of autoimmune diseases. One study showed association of both rheumatoid arthritis and T1DM with specific polymorphisms (*IL2RA*-rs2104286 and *PTPN22*-rs6679677; TABLE 2)[18]. T1DM, rheumatoid arthritis and IBD also show association with MHC alleles[74–76]. These findings suggest common underlying aetiological pathways (and therapeutic targets) for several, common autoimmune disorders[77].

Cancer. GWA studies of cancer based on common, inherited SNPs are useful for the identification of germline risk alleles, but not somatic mutations. Three GWA studies sought inherited association signals in prostate cancer[53,79,79]; FIG. 2 shows details of the discovery phase of one of these studies. An association signal at chromosome 8q24 that had previously been identified by linkage analysis[80] was replicated in two GWA studies[53,79]. In addition, these studies identified a second 8q24 association, approximately 300 kb upstream from the first. As yet, the functional basis of these associations is unclear. Although individual 8q24 alleles showed modest estimated genetic effects, the cumulative effect of several loci fit a multiplicative model that conferred a population-attributable risk (PAR), that is, an expected reduction in prostate-cancer incidence if the risk alleles did not exist in the population, of up to 68%[81]. As noted above, PAR values are strongly affected by allele frequency and represent only an approximate measure of the contribution of those alleles to disease incidence.

One study of prostate cancer[79] identified a *TCF2* (also known as *HNF1B*) susceptibility allele. Intriguingly, this allele appeared to diminish the risk of T2DM (TABLE 2),

Family-based association study
Evaluation of the frequency of co-transmission of an allele and a trait from parents to offspring. Co-transmission of an allele and trait to offspring more often than expected by chance indicates that the allele might change the likelihood of the trait.

Table 3 | **Loci and variants exhibiting association with type 2 diabetes mellitus in GWA studies**

T2DM phenotype	Locus	Variants (rs)	Refs
Susceptibility	TCF2	4430796, 7501939	78
	TCF7L2	4506565, 7903146, 7901695	18,54–57,63
	PPARG	1801282	55,63
	KCNJ11	5219, 5212	54,55,63
	SLC30A8	13266634, rs118253964	55,56,63
	HHEX	1111875	55,63
	IGF2BP2	4402960	54,55,63
	CDKAL1	9456871, 7754840, 10946398, 7756992	18,55,57,63
	CDKN2A/B	10811661, 564398	54,55,63
	Chromosome 11, intergenic	9300039	55
	FTO	9939609, 7193144, 8050136	18,55,63
Low-density lipoprotein	APOE	4420638	54
High-density lipoprotein	CETP	1800775	54
High-triglyceride level	LPL	17482753	54
	GCKR	780094	54

Above variants are associated $P < 5 \times 10^{-7}$. T2DM, type 2 diabetes mellitus.

possibly representing antagonistic pleiotropy. This is supported by epidemiological evidence which suggests that diabetic men have a slightly lower prostate cancer risk than non-diabetic men[82].

Another allele exhibiting association in two diseases is rs6983267 at chromosome 8q24, which has shown replicated associations with prostate and colorectal cancer[79,82–84] (TABLE 2).

Three GWA studies sought inherited associations with breast cancer[19,85,86]. Although each study identified significant novel loci, two genes and one allele were each supported in two studies.

Complex traits

In addition to common diseases, GWA studies are applicable to complex traits. One study undertook GWA with numerous quantitative and categorical memory-associated endophenotypes[87]. Despite a small discovery cohort (341 individuals), associations with the *KIBRA* (also known as *WWC1*) gene have been replicated[87–89]. A notable innovation in this study was that associations were sought with multi-scale and multi-modality endophenotypes; that is, performance in seven memory-associated tests and functional magnetic resonance image-based measures of the hippocampus during three memory-associated tests. This study provides evidence that progress can be made in the elucidation of the genetic determinants of subjective, qualitative neurologic traits by using objective, quantitative, surrogate endophenotypes.

As well as identifying novel associations, GWA studies have confirmed several susceptibility genes that were previously established by linkage analysis in large pedigrees. For example, a GWA study of late-onset Alzheimer's disease (LOAD) identified the well-established *APOE*-susceptibility allele[90]. This association was also replicated in a study that genotyped 17,343 putative functional cSNPs[23].

Antagonistic pleiotropy
A single gene that influences multiple competing phenotypes such that beneficial effects of a trait created by the gene are offset by losses in other traits.

A remaining problem with large GWA studies is the cost of genotyping, but one study provided evidence that sample pooling strategies might help to overcome this issue. In a GWA study of bipolar disorder, investigators created 39 pools, containing DNA from 2,672 individuals[91]. These pools were used for both discovery and replication experiments. Pools were individually genotyped for 555,235 SNPs and normalized allele frequencies were inferred from intensity data. Replicates were assayed for each pool. Thirty-seven SNPs showing allele frequency differences in both cohorts were individually genotyped and one SNP retained a significant association. The aforementioned WTCCC study also studied bipolar disorder, identifying an association at 16p12 (REF. 18). One locus, for glutamate receptor, metabotropic 7 (*GRM7*), showed association in both studies.

The rate of publication of GWA studies continues to increase. Recent studies have investigated asthma[92], nicotine dependence[93], coronary artery disease[19,26], atrial fibrillation[94], prolonged QT interval and sudden cardiac death[95,96], coeliac disease[97], lung cancer[98], psoriasis[21] and liver cirrhosis[25], among others (TABLE 1).

Initial conclusions on the utility of GWA

The utility of GWA studies for the identification of novel genomic associations with complex diseases has unambiguously been established over the past year. In general, GWA studies have employed large case–control cohorts featuring both familial and sporadic cases, categorical trait definitions and up to half a million commonly polymorphic SNPs. To date, with the exception of *CFH* in AMD, the estimated genetic effects of replicated associations have been uniformly and surprisingly small.

Encouragingly, most associated haplotype intervals identified to date are sufficiently small to feature a single gene. In large measure, this reflects the use of several,

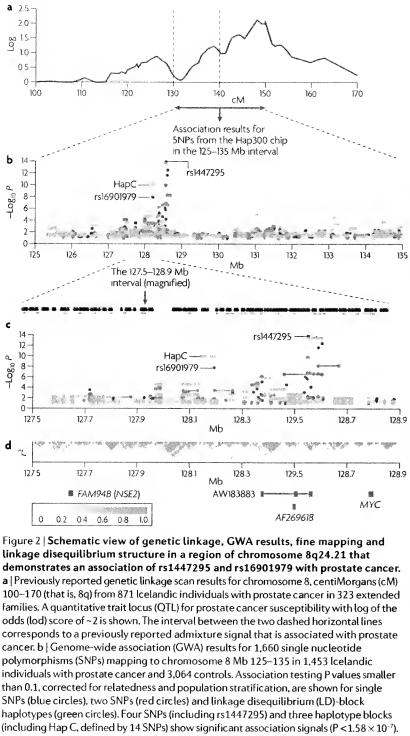


Figure 2 | **Schematic view of genetic linkage, GWA results, fine mapping and linkage disequilibrium structure in a region of chromosome 8q24.21 that demonstrates an association of rs1447295 and rs16901979 with prostate cancer.** **a** | Previously reported genetic linkage scan results for chromosome 8, centiMorgans (cM) 100–170 (that is, 8q) from 871 Icelandic individuals with prostate cancer in 323 extended families. A quantitative trait locus (QTL) for prostate cancer susceptibility with log of the odds (lod) score of ~2 is shown. The interval between the two dashed horizontal lines corresponds to a previously reported admixture signal that is associated with prostate cancer. **b** | Genome-wide association (GWA) results for 1,660 single nucleotide polymorphisms (SNPs) mapping to chromosome 8 Mb 125–135 in 1,453 Icelandic individuals with prostate cancer and 3,064 controls. Association testing P values smaller than 0.1, corrected for relatedness and population stratification, are shown for single SNPs (blue circles), two SNPs (red circles) and linkage disequilibrium (LD)-block haplotypes (green circles). Four SNPs (including rs1447295) and three haplotype blocks (including HapC, defined by 14 SNPs) show significant association signals ($P < 1.58 \times 10^{-7}$). Single SNP association two-sided P values were derived using Fisher's exact test and were unadjusted for multiple comparisons. Association testing of haplotype block P values were carried out using the expectation-maximization (EM) algorithm directly for the observed data. **c** | Association results from **b**, shown in greater detail, for a 1.4 Mb interval on 8q24.21. Filled black circles represent 225 SNPs and the orange boxes represent recombination hotspots (calculated from the HapMap using the likelihood ratio test). **d** | LD between SNPs, measured by the square of the correlation coefficient calculated for each pairwise comparison of SNPs (r^2) from the Centre d'Etude du Polymorphisme Humain from Utah (CEU) HapMap population for the 225 SNPs in **c**; the blue boxes at the bottom indicate the location of the *FAM84B*, *AF268618* and *MYC* genes and the AW183883 expressed sequence tag. Figure modified, with permission, from *Nature Genetics* REF 55 © 2007 Macmillan Publishers Ltd.

outbred populations in confirmatory, fine-mapping studies. Even when the association is within a single gene, the predisposing variant might affect an adjacent gene, as in adult lactose intolerance[90]. Although some association intervals have been found to contain a single, unequivocally functional gene variant, the causality of alleles has been established in only a minority of cases. Causal alleles identified to date do not yet show much difference in genetic mechanism from those identified in Mendelian disorders; this could reflect ascertainment bias[1,83,91].

Many genes identified by GWA were not candidate genes previously, highlighting the hypothesis-informing value of genetic studies. Already, there are examples of potentially tractable therapeutic targets that had not previously been considered in a disease or trait. As yet, the confluence of associated genes into biological networks and pathways is at an early stage. In part, this reflects scant or incorrect annotation of many genes. There appears to be a significant conservation of associations of common alleles between human populations. Thus, to date, it appears that GWA studies are fulfilling expectations with regard to the elucidation of molecular mechanisms underpinning poorly understood, common diseases.

In the few informative studies reported to date, endophenotypes have been highly instructive in dissecting the network or pathway that is perturbed by an individual allele, which affects a complex trait. It is particularly exciting to see the application of multi-mode endophenotypes, such as combinations of psychological testing, brain imaging and gene expression[87]. This is clearly an area of potential opportunity.

The cost of enrolling the very large cohorts that are needed to discover and validate alleles with small effect sizes has hitherto precluded the collection and integration of rich, accurate clinical metadata. It is likely that future studies will use a much greater stratification of traits than the phenotypically crude studies reported so far. Recent GWA studies of breast cancer provide a good example of the added genetic complexity that can be revealed by trait stratification[49,85,86]. In addition, following replication of associations with categorical traits, it is anticipated that targeted genotypic examination of many endophenotypes will be highly instructive in the dissection of the role of individual alleles in disease pathogenesis.

GWA studies show significant potential to redefine disease classification. In some cases, GWA studies are identifying molecular factors that enable patient stratification and might prove useful in personalized medicine. Cancers provide the clearest examples of this to date. In other cases, exemplified by IBD, GWA studies are pointing to common molecular underpinnings in diseases that were believed to be distinct. In restless leg syndrome, replicated associations have provided concrete evidence that the phenotype represents a bona fide neurological disorder[100,101]. In mental illness, there is great anticipation that GWA studies will provide an objective, molecular revision of disease categorization.

Many questions remain concerning the genetic architecture of common diseases. These include the extent of locus and allelic heterogeneity, fit with an additive-threshold model or, alternatively, the extent of epistasis

(the relative contributions of rare and common, and high and low, penetrance alleles) and various types of variation, from genome rearrangements to SNPs. GWA studies are not designed to evaluate these questions. Once loci have been identified, however, methods such as deep resequencing can nominate candidate susceptibility alleles and provide data for the evaluation of genetic architecture[102]. Meaningful, individual risk determinations will require the identification of causal alleles, the development of multiplexed molecular diagnostics and significant modelling.

Future developments and implications

The trends observed in recent GWA studies are anticipated to continue. Chips with 900,000 and 1,000,000 million SNPs were recently launched and genotyping accuracies have improved. Cohort sizes are steadily increasing and biobanks of unparalleled size and phenotype definition are being established. Combinations of genotype- and haplotype-based associations are becoming more prevalent. Experimental designs and statistical methods are also becoming more uniform, enabling more meaningful meta-analysis. In particular, the emergence of adaptive designs and the use of Bayesian inferential methods will produce a probabilistic synthesis from combined analyses[65]. Importantly, this will provide an intuitive framework for combining information from multiple studies, resulting in more effective detection and replication of weak associations[103].

As noted above, phenotypes studied to date have been crude. The use of endophenotypes is expected to increase significantly. In particular, biomarker phenotypes are anticipated to become widely used. These will probably include gene expression, proteomic, metabolomic and imaging biomarkers. As determinants of complex traits are identified, genetic stratification will become possible, potentially reducing the genetic complexity of traits and enabling the identification of additional association signals. An example of this was the recent use of periodic limb movements and serum ferritin levels in GWA studies of restless leg syndrome[100]. An area of substantial future interest for the pharmaceutical industry will be pharmacogenetic GWA studies to identify markers for patient stratification in clinical trials. Comprehensive pharmacogenetic information will, in turn, facilitate the practice of personalized medicine. Pharmacogenetic GWA studies and early adoption of personalized therapy are likely to be used in the selection of expensive or chronic medications in life threatening conditions or where the therapeutic index is narrow or adverse event concerns are high, such as cancer chemotherapy.

Despite the current excitement, GWA studies have only been able to account for a small proportion of the expected genetic variance in complex traits[58,102]. This is not surprising given current limitations. First, current GWA studies are designed to identify common risk alleles that are predicted to be important in complex disorders under the common disease/common alleles hypothesis[58]. Increasing evidence suggests that some complex disorders and traits, such as schizophrenia, hypercholesterolaemia and body mass, are genetically heterogeneous[106, 107]. The genetic basis of such diseases is more likely to conform to the common trait/rare variant hypothesis, which proposes that many rare variants exist, with substantial allelic heterogeneity at causal loci[58,108,109]. The GWA approach is unable to detect susceptibility loci that harbour numerous, individually rare (recent), polymorphisms. Instead, a resequencing approach will be needed to identify rare alleles. Encouragingly, massively parallel sequencing methods provide a potential solution[102,107,110], suggesting disease-specific rare alleles and recent mutations that provide supplementary genotyping array content. Second, a proportion of the genome cannot effectively be examined on the basis of tag SNP genotypes. Approximately 20% of the genome is comprised of recombination hotspots that are not amenable to LD-based approaches[7]. Alternatively, at recombination coldspots, haplotype blocks might be too large for unambiguous identification of causal loci. The extent of the effect of genomic copy number variation (CNV) on association signals is not yet clear, although recent genotyping arrays do provide CNV information. Insufficient numbers of cases will be available for GWA studies of many orphan diseases, uncommon disease complications or adverse events. For some common diseases, these considerations could obfuscate a substantial proportion of the genetic variance. Supplementation of genotyping array content reflective of CNV regions should, however, circumvent some of these limitations. Use of adaptive statistical methods and resampling strategies might also circumvent the need for thousands of affected individuals in studies of orphan diseases[63].

GWA successes are creating substantial need for downstream genetics, biochemistry and cell biology efforts to confirm the biological relevance of genotype–phenotype associations and to elucidate the underlying mechanisms of disease. This is especially true of association signals in gene deserts or alleles without apparent functional consequence. Translation of the fruits of GWA studies to clinical practice will require the derivation of predictive models of the genetic architecture of complex traits that evaluate with much greater precision the contributions of factors such as epistasis, genocopies, phenocopies and penetrance.

1. Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature Genet.* **33** (Suppl.), 228–237 (2003).
2. Freimer, N. & Sabatti, C. The use of pedigree, sib-pair and association studies of common diseases for genetic mapping and epidemiology. *Nature Genet.* **36**, 1045–1051 (2004).
3. Goring, H. H., Terwilliger, J. D. & Blangero, J. Large upward bias in estimation of locus-specific effects from genomewide scans. *Am. J. Hum. Genet.* **69**, 1357–1369 (2001).
4. Lander, E. & Kruglyak, L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genet.* **11**, 241–247 (1995).
5. Chakravarti, A. Population genetics — making sense out of sequence. *Nature Genet.* **21**, 56–60 (1999).
6. Reich, D. E. & Lander, E. S. On the allelic spectrum of human disease. *Trends Genet.* **17**, 502–510 (2001).
7. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
8. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
9. Hirschhorn, J. N., Lohmueller, K., Byrne, E. & Hirschhorn, K. A comprehensive review of genetic association studies. *Genet. Med.* **4**, 45–61 (2002).
10. Ioannidis, J. P., Ntzani, E. E., Trikalinos, T. A. & Contopoulos-Ioannidis, D. G. Replication validity of genetic association studies. *Nature Genet.* **29**, 306–309 (2001).
11. Cardon, L. R. & Bell, J. I. Association study designs for complex diseases. *Nature Rev. Genet.* **2**, 91–99 (2001).

12 Cardon, L. R. & Palmer, L. J. Population stratification and spurious allelic association. *Lancet* **361**, 598–604 (2003)
13 Redden, D. T. & Allison, D. B. Nonreplication in genetic association studies of obesity and diabetes research. *J. Nutr.* **133**, 3525–3526 (2003)
14 Sillanpaa, M. J. & Auranen, K. Replication in genetic studies of complex traits. *Ann. Hum. Genet.* **68**, 646–657 (2004)
15 Lohmueller, K. E., Pearce, C. L., Pike, M., Lander, E. S. & Hirschhorn, J. N. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nature Genet.* **33**, 177–182 (2003).
16 Risch, N. J. Searching for genetic determinants in the new millennium. *Nature* **405**, 847–856 (2000)
17 Chanock, S. J. *et al.* Replicating genotype–phenotype associations. *Nature* **447**, 655–660 (2007).
18 Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
The largest GWA study undertaken to date.
19 Hunter, D. J. *et al.* A genome-wide association study identifies alleles in *FGFR2* associated with risk of sporadic postmenopausal breast cancer. *Nature Genet.* **39**, 870–874 (2007).
20 Ozaki, K. *et al.* Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. *Nature Genet.* **32**, 650–654 (2002).
The first large scale association study of a complex human disorder.
21 Cargill, M. *et al.* A large-scale genetic association study confirms *IL18* and leads to the identification of *IL23R* as psoriasis-risk genes. *Am. J. Hum. Genet.* **80**, 273–290 (2007)
22 Clark, A. G. & Li, J. Conjuring SNPs to detect associations. *Nature Genet.* **39**, 815–816 (2007).
23 Grupe, A. *et al.* Evidence for novel susceptibility genes for late-onset Alzheimer's disease from a genome-wide association study of putative functional variants. *Hum. Mol. Genet.* **16**, 865–875 (2007).
24 Hampe, J. *et al.* A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in *ATG16L1*. *Nature Genet.* **39**, 207–211 (2007).
25 Huang, H. *et al.* Identification of two gene variants associated with risk of advanced cervical fibrosis in patients with chronic hepatitis C. *Gastroenterology* **130**, 1679–1687 (2006).
26 Luke, M. M. *et al.* A polymorphism in the protease-like domain of apolipoprotein(a) is associated with severe coronary artery disease. *Arterioscler. Thromb. Vasc. Biol.* **27**, 2030–2036 (2007).
27 Shiffman, D. *et al.* Identification of four gene variants associated with myocardial infarction. *Am. J. Hum. Genet.* **77**, 596–605 (2005).
28 Smyth, D. J. *et al.* A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (*IFIH1*) region. *Nature Genet.* **38**, 617–619 (2006).
29 Clarke, R. *et al.* Lymphotoxin-α gene and risk of myocardial infarction in 6,928 cases and 2,712 controls in the ISIS case-control study. *PLoS Genet.* **2**, e107 (2006)
30 Kimura, A. *et al.* Lack of association between *LTA* and *LGALS2* polymorphisms and myocardial infarction in Japanese and Korean populations. *Tissue Antigens* **69**, 265–269 (2007).
31 Koch, W. *et al.* Association of variants in the *BAT1–NFKBIL1–LTA* genomic region with protection against myocardial infarction in Europeans. *Hum. Mol Genet.* **16**, 1821–1827 (2007).
32 Laxton, R., Pearce, E., Kyriakou, T. & Ye, S. Association of the lymphotoxin-α gene Thr26Asn polymorphism with severity of coronary atherosclerosis. *Genes Immun.* **6**, 539–541 (2005)
33 Mizuno, H. *et al.* Impact of atherosclerosis-related gene polymorphisms on mortality and recurrent events after myocardial infarction. *Atherosclerosis* **185**, 400–405 (2006)
34 Sedlacek, K. *et al.* Lymphotoxin-α and galectin-2 SNPs are not associated with myocardial infarction in two different German populations. *J. Mol. Med.* **85**, 997–1004 (2007).
35 Yamada, A. *et al.* Lack of association of polymorphisms of the lymphotoxin α gene with myocardial infarction in Japanese. *J. Mol. Med.* **82**, 477–483 (2004).
36 Klein, R. J. *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385–389 (2005).
Discovery of a single variant that explains a large component of the genetic variance in a common human disease.
37 Hageman, G. S. *et al.* A common haplotype in the complement regulatory gene factor H (*HF1/CFH*) predisposes individuals to age-related macular degeneration. *Proc. Natl Acad. Sci. USA* **102**, 7227–7232 (2005)
38 Magnusson, K. P. *et al.* CFH Y402H confers similar risk of soft drusen and both forms of advanced AMD. *PLoS Med.* **3**, e5 (2006).
39 Souied, E. H. *et al.* Y402H complement factor H polymorphism associated with exudative age-related macular degeneration in the French population. *Mol. Vis.* **11**, 1135–1140 (2005)
40 Zareparsi, S. *et al.* Strong association of the Y402H variant in complement factor H at 1q32 with susceptibility to age-related macular degeneration. *Am. J. Hum. Genet.* **77**, 149–153 (2005)
41 Gold, B. *et al.* Variation in factor B (*BF*) and complement component 2 (*C2*) genes is associated with age-related macular degeneration. *Nature Genet.* **38**, 458–462 (2006).
42 Yates, J. R. *et al.* Complement C3 variant and the risk of age-related macular degeneration. *N. Engl. J. Med.* **357**, 553–561 (2007).
43 Duerr, R. H. *et al.* A genome-wide association study identifies *IL23R* as an inflammatory bowel disease gene. *Science* **314**, 1461–1463 (2006).
44 Libioulle, C. *et al.* Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of *PTGER4*. *PLoS Genet.* **3**, e58 (2007).
45 Rioux, J. D. *et al.* Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nature Genet.* **39**, 596–604 (2007).
46 Parkes, M. *et al.* Sequence variants in the autophagy gene *IRGM* and multiple other replicating loci contribute to Crohn's disease susceptibility. *Nature Genet.* **39**, 830–832 (2007).
47 Yamazaki, K. *et al.* Association analysis of genetic variants in *IL23R*, *ATG16L1* and 5p13.1 loci with Crohn's disease in Japanese patients. *J. Hum. Genet* **52**, 575–583 (2007).
48 Hugot, J. P. *et al.* Association of *NOD2* leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**, 599–603 (2001).
49 Ogura, Y. *et al.* A frameshift mutation in *NOD2* associated with susceptibility to Crohn's disease. *Nature* **411**, 603–606 (2001).
50 Xavier, R. J. & Podolsky, D. K. Unravelling the pathogenesis of inflammatory bowel disease. *Nature* **448**, 427–434 (2007).
51 Todd, J. A. *et al.* Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nature Genet.* **39**, 857–864 (2007).
52 Raelson, J. V. *et al.* Genome-wide association study for Crohn's disease in the Quebec Founder Population identifies multiple validated disease loci. *Proc. Natl Acad. Sci. USA* **104**, 14747–14752 (2007).
53 Gudmundsson, J. *et al.* Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nature Genet.* **39**, 631–637 (2007).
54 Saxena, R. *et al.* Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316**, 1331–1336 (2007).
55 Scott, L. J. *et al.* A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* **316**, 1341–1345 (2007).
56 Sladek, R. *et al.* A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881–885 (2007).
57 Steinthorsdottir, V. *et al.* A variant in *CDKAL1* influences insulin response and risk of type 2 diabetes. *Nature Genet.* **39**, 770–775 (2007).
58 Liu, P. Y. *et al.* A survey of haplotype variants at several disease candidate genes: the importance of rare variants for complex diseases. *J. Med. Genet.* **42**, 221–227 (2005).
59 Morris, R. W. & Kaplan, N. L. On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. *Genet. Epidemiol.* **23**, 221–233 (2002).
60 Zhang, K., Calabrese, P., Nordborg, M. & Sun, F. Haplotype block structure and its applications to association studies: power and study designs. *Am J. Hum. Genet.* **71**, 1386–1394 (2002)
61 Zhang, K. & Sun, F. Assessing the power of tag SNPs in the mapping of quantitative trait loci (QTL) with extremal and random samples. *BMC Genet.* **6**, S1 (2005)
62 Drysdale, C. M. *et al.* Complex promoter and coding region β2-adrenergic receptor haplotypes alter receptor expression and predict in vivo responsiveness. *Proc. Natl Acad. Sci. USA* **97**, 10483–10488 (2000)
63 Zeggini, E. *et al.* Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* **316**, 1336–1341 (2007)
64 Grant, S. F. *et al.* Variant of transcription factor 7-like 2 (*TCF7L2*) gene confers risk of type 2 diabetes. *Nature Genet.* **38**, 320–323 (2006).
65 Yi, F., Balsbaugh, P. L. & Shi, T. TCF4 mediates cell type-specific regulation of proglucagon gene expression by β-catenin and glycogen synthase kinase-3β. *J. Biol. Chem.* **280**, 1457–1464 (2005)
66 Pearson, E. R. *et al.* Variation in *TCF7L2* influences therapeutic response to sulfonylureas: a GoDARTs study. *Diabetes* **56**, 2178–2182 (2007).
67 Florez, J. C. *et al.* Haplotype structure and genotype–phenotype correlations of the sulfonylurea receptor and the islet ATP-sensitive potassium channel gene region. *Diabetes* **53**, 1360–1368 (2004)
68 Helgason, A. *et al.* Refining the impact of *TCF7L2* gene variants on type 2 diabetes and adaptive evolution. *Nature Genet.* **39**, 218–225 (2007).
69 Weedon, M. N. *et al.* Combining information from common type 2 diabetes risk polymorphisms improves disease prediction. *PLoS Med.* **3**, e374 (2006).
70 Stephens, J. C. *et al.* Haplotype variation and linkage disequilibrium in 313 human genes. *Science* **293**, 489–493 (2001).
71 Frayling, T. M. *et al.* A common variant in the *FTO* gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* **316**, 889–894 (2007).
72 Dina, C. *et al.* Variation in *FTO* contributes to childhood obesity and severe adult obesity. *Nature Genet.* **39**, 724–726 (2007)
73 Rich, S. S. *et al.* The Type 1 Diabetes Genetics Consortium. *Ann. NY Acad. Sci.* **1079**, 1–8 (2006).
74 Ahmad, T., Marshall, S. E. & Jewell, D. Genetics of inflammatory bowel disease: the role of the HLA complex. *World J. Gastroenterol.* **12**, 3628–3635 (2006).
75 Orozco, G., Rueda, B. & Martin, J. Genetic basis of rheumatoid arthritis. *Biomed. Pharmacother.* **60**, 656–662 (2006).
76 Sia, C. & Weinem, M. The role of HLA class I gene variation in autoimmune diabetes. *Rev. Diabet. Stud.* **2**, 97–109 (2005).
77 Criswell, L. A. *et al.* Analysis of families in the multiple autoimmune disease genetics consortium (MADGC) collection: the *PTPN22* 620W allele associates with multiple autoimmune phenotypes. *Am. J. Hum. Genet.* **76**, 561–571 (2005).
78 Gudmundsson, J. *et al.* Two variants on chromosome 17 confer prostate cancer risk, and the one in *TCF2* protects against type 2 diabetes. *Nature Genet.* **39**, 977–983 (2007).
79 Yeager, M. *et al.* Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nature Genet.* **39**, 645–649 (2007).
80 Amundadottir, L. T. *et al.* A common variant associated with prostate cancer in European and African populations. *Nature Genet.* **38**, 652–658 (2006).
81 Haiman, C. A. *et al.* Multiple regions within 8q24 independently affect risk for prostate cancer. *Nature Genet.* **39**, 638–644 (2007).
82 Rodriguez, C. *et al.* Diabetes and risk of prostate cancer in a prospective cohort of US men. *Am J. Epidemiol.* **161**, 147–152 (2005)
83 Knight, J. C. Regulatory polymorphisms underlying complex disease traits. *J. Mol. Med.* **83**, 97–109 (2005).
84 Thomas, P. D. & Kejariwal, A. Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: evolutionary evidence for differences in molecular effects. *Proc. Natl Acad. Sci. USA* **101**, 15398–15403 (2004).
85 Easton, D. F. *et al.* Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447**, 1087–1093 (2007)
86 Stacey, S. N. *et al.* Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. *Nature Genet.* **39**, 865–869 (2007).

87. Papassotiropoulos, A. *et al.* Common Kibra alleles are associated with human memory performance. *Science* **314**, 475–478 (2006).
88. Rodriguez-Rodriguez, E. *et al.* Age-dependent association of *KIBRA* genetic variation and Alzheimer's disease risk. *Neurobiol. Aging* 16 Aug 2007 (doi:10.1016/j.neurobiolaging.2007.07.005).
89. Schaper, K., Kolsch, H., Popp, J., Wagner, M. & Jessen, F. *KIBRA* gene variants are associated with episodic memory in healthy elderly. *Neurobiol. Aging* 10 Mar 2007 (doi:10.1016/j.neurobiolaging.2007.02.001).
90. Coon, K. D. *et al.* A high-density whole genome association study reveals that *APOE* is the major susceptibility gene for sporadic late-onset Alzheimer's disease. *J. Clin. Psychiatry* **68**, 613–618 (2007).
91. Baum, A. E. *et al.* A genome-wide association study implicates diacylglycerol kinase eta (*DGKH*) and several other genes in the etiology of bipolar disorder. *Mol. Psychiatry* 8 May 2007 (doi:10.1038/sj.mp.4002012).
92. Moffatt, M. F. *et al.* Genetic variants regulating *ORMDL3* expression contribute to the risk of childhood asthma. *Nature* **448**, 470–473 (2007).
93. Bierut, L. J. *et al.* Novel genes identified in a high-density genome wide association study for nicotine dependence. *Hum. Mol. Genet.* **16**, 24–35 (2007).
94. Gudbjartsson, D. F. *et al.* Variants conferring risk of atrial fibrillation on chromosome 4q25. *Nature* **448**, 353–357 (2007).
95. Aarnoudse, A. J. *et al.* Common *NOS1AP* variants are associated with a prolonged QTc interval in the Rotterdam Study. *Circulation* **116**, 10–16 (2007).
96. Arking, D. E. *et al.* A common genetic variant in the *NOS1* regulator *NOS1AP* modulates cardiac repolarization. *Nature Genet.* **39**, 644–651 (2006).
97. van Heel, D. A. *et al.* A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. *Nature Genet.* **39**, 827–829 (2007).
98. Spinola, M. *et al.* Genome-wide single nucleotide polymorphism analysis of lung cancer risk detects the *KLF6* gene. *Cancer Lett.* **251**, 311–316 (2007).
99. Olds, L. C. & Sibley, E. Lactase persistence DNA variant enhances lactase promoter activity in vitro: functional role as a cis regulatory element. *Hum. Mol. Genet.* **12**, 2333–2340 (2003).
100. Stefansson, H. *et al.* A genetic risk factor for periodic limb movements in sleep. *N. Engl. J. Med.* **357**, 639–647 (2007).
101. Winkelmann, J. *et al.* Genome-wide association study of restless legs syndrome identifies common variants in three genomic regions. *Nature Genet.* **39**, 1000–1006 (2007).
102. Altshuler, D. & Daly, M. Guilt beyond a reasonable doubt. *Nature Genet.* **39**, 813–815 (2007).
103. Hunter, D. J. & Kraft, P. Drinking from the fire hose — statistical issues in genomewide association studies. *N. Engl. J. Med.* **357**, 436–439 (2007).
104. Ahituv, N. *et al.* Medical sequencing at the extremes of human body mass. *Am. J. Hum. Genet.* **80**, 779–791 (2007).
105. Cohen, J. C. *et al.* Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* **305**, 869–872 (2004).
An example of a phenotype that fits the common disorder: rare alleles hypothesis.
106. Fanous, A. H. & Kendler, K. S. Genetic heterogeneity, modifier genes, and quantitative phenotypes in psychiatric illness: searching for a framework. *Mol. Psychiatry* **10**, 6–13 (2005).
107. McClellan, J. M., Susser, E. & King, M. C. Schizophrenia: a common disease caused by multiple rare alleles. *Br. J. Psychiatry* **190**, 194–199 (2007).
108. Pritchard, J. K. Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.* **69**, 124–137 (2001).
109. Pritchard, J. K. & Cox, N. J. The allelic architecture of human disease genes: common disease–common variant...or not? *Hum. Mol. Genet.* **11**, 2417–2423 (2002).
110. Kryukov, G. V., Pennacchio, L. A. & Sunyaev, S. R. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am. J. Hum. Genet.* **80**, 727–739 (2007).
111. Tomlinson, I. *et al.* A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nature Genet.* **39**, 984–988 (2007).
112. Zanke, B. W. *et al.* Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nature Genet.* **39**, 989–994 (2007).
113. Haiman, C. A. *et al.* A common genetic risk factor for colorectal and prostate cancer. *Nature Genet.* **39**, 954–956 (2007).

Acknowledgements

A Deo lumen, ab omne auxilium. This work was partially supported by National Institutes of Health grants N01A000,064 and U01A106,569, and by National Science Foundation grant 0524,775. The authors thank the reviewers for their helpful suggestions.

DATABASES

Entrez Gene: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene
APOE | ATG16L1 | BAD | C3 | CFB | CFH | CLEC16A | FID | GRM7 | HLA-B27 | IL23R | IRF1 | KLRA1 | LCE3B | NOD2 | PTGER4 | PTPN1 | PTPN22 | RELA | TCF7L2 | WWC1

OMIM: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM
age-related macular degeneration | acute myocardial infarction | Alzheimer's disease | ankylosing spondylitis | breast cancer | colorectal cancer | Crohn's disease | prostate cancer | restless legs syndrome | type 1 diabetes mellitus | type 2 diabetes mellitus | ulcerative colitis

FURTHER INFORMATION

Stephen Kingsmore's homepage: http://www.cgsmy

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

Frayling, TM, Nature Reviews Genet 8:657-662 (2007)

Genome-wide association studies provide new insights into type 2 diabetes aetiology

Timothy M. Frayling

Abstract | Human geneticists are currently in the middle of a race. Thanks to a new technology in the form of 'genome-wide chips', investigators can potentially find many novel disease genes in one large experiment. Type 2 diabetes has been hot out of the blocks with six recent publications that together provide convincing evidence for six new gene regions involved in the condition. Together with candidate approaches, these studies have identified 11 confirmed genomic regions that alter the risk of type 2 diabetes in the European population. One of these regions, the fat mass and obesity associated gene (*FTO*), represents by far the best example of an association between common variation and fat mass in the general population.

Genome-wide association studies (GWAS) promised to greatly enhance our understanding of the genetic basis of common and complex diseases. Companies such as Affymetrix and Illumina have developed chips that can capture information from more than two-thirds of the common variation in the human genome. Approximately 300–500,000 SNPs can be analysed using these chips. Importantly, this can now be done for several thousands of DNA samples at costs that are within the scope of large project grants.

This technology recently facilitated rapid progress in type 2 diabetes genetic research. This is all the more remarkable because type 2 diabetes does not have a strong genetic component compared with some other common traits, and was previously described as 'a geneticist's nightmare'[1,2]. Nevertheless, early results have been excellent, yielding six new replicating gene regions.

Here I discuss the insights into type 2 diabetes genetics that have been provided by these new findings. I consider where diabetes genetic studies might go from here, and present a perspective that may be applicable to other common traits. I also briefly discuss the wider implications that surround the identification of a common gene that predisposes to type 2 diabetes by altering fat mass.

The geneticist's nightmare

Type 2 diabetes is one of the leading health problems throughout the developed world and is becoming increasingly important in the developing world. It has risen in prevalence dramatically in the past two generations as we have come to lead increasingly sedentary lifestyles and as food has become more plentiful. Obesity, defined as a body mass index (BMI) of greater than 30 kgm^{-2} increases the risk of type 2 diabetes. Such a strong environmental component to a disease should perhaps have deterred geneticists from studying the disorder. However, there are many obese people who do not suffer from diabetes and many non-obese people who do, showing that obesity is not the only factor involved in the aetiology of type 2 diabetes (FIG. 1).

In the past 10 years, geneticists have devoted a large amount of effort to finding type 2 diabetes genes. These efforts have included many candidate-gene studies, extensive efforts to fine map linkage signals[3], and an international linkage consortium that was perhaps the best example of a multi-centre collaboration in common-disease genetics. Of these efforts, only the candidate-gene studies produced unequivocal evidence for common variants involved in type 2 diabetes. These are the E23K variant in the potassium inwardly-rectifying channel, subfamily J, member 11 (*KCNJ11*) gene[4–5], the P12A variant in the peroxisome proliferator-activated receptor-γ (*PPARG*) gene[6], and common variation in the transcription factor 2, hepatic (*TCF2*)[6,9] and the Wolfram syndrome 1 (*WFS1*)[10] genes. All of these genes encode proteins that have strong biological links to diabetes. Rare, severe mutations in all four cause monogenic forms of diabetes[11–14], and two are targets of anti-diabetic therapies: *KCNJ11* encodes a component of a potassium channel with a

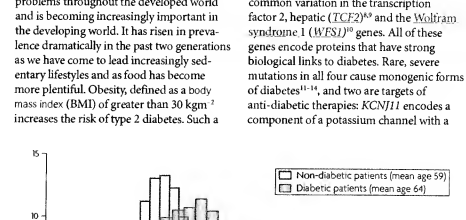


Figure 1 | **Distribution of body mass index in type 2 diabetic patients compared with non-diabetic individuals of a similar age.** Individuals come from the Diabetes Audit and Research in Tayside (DARTs) study in Scotland[20].

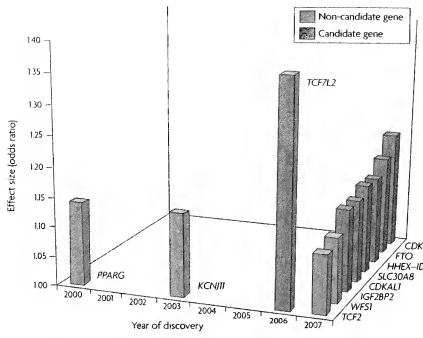


Figure 2 | **Effect sizes of the 11 common variants confirmed to be involved in type 2 diabetes risk.** The x axis gives the year that published evidence reached the levels of statistical confidence that are now accepted as necessary for genetic association studies. *CDKAL1*, CDK5 regulatory subunit-associated protein 1-like 1; *CDKN2*, cyclin-dependent kinase inhibitor 2A; *FTO*, fat mass and obesity-associated; *HHEX*, haematopoietically expressed homeobox; *IDE*, insulin-degrading enzyme; *IGF2BP2*, insulin-like growth factor 2 mRNA-binding protein 2; *KCNJ11*, potassium inwardly-rectifying channel, subfamily J, member 11; *PPARG*, peroxisome proliferator-activated receptor-γ gene; *SLC30A8*, solute carrier family 30 (zinc transporter), member 8; *TCF2*, transcription factor 2, hepatic; *TCF7L2*, transcription factor 7-like 2 (T-cell specific, HMG-box); *WFS1*, Wolfram syndrome 1.

key role in β-cell physiology that is a target for the sulphonylurea class of drugs, and *PPARG* encodes a transcription factor involved in adipocyte differentiation that is a target for the thiazolodinedione class of drugs.

In 2006, deCODE genetics identified common variation in the *TCF7L2* gene as a type 2 diabetes risk gene region[15]. This result was encouraging for two reasons. First, this study analysed more than 200 markers across a region of linkage, but the variants that were found to alter risk did not explain the linkage signal, suggesting that a non-candidate-gene or region-based association effort (such as a GWAS) could work. Second, *TCF7L2* was a completely unexpected gene — this showed that a genome-wide approach could uncover previously unexpected disease pathways.

In early 2007, GWAS provided by far the biggest increment to date in our knowledge of the genetics of this common health problem.

Six new gene regions identified

Together, the six recent GWAS papers provide convincing evidence for six new gene regions involved in type 2 diabetes[16–21]; a seventh publication describes how one of these variants alters BMI and represents by far the best example of an association between common genetic variation and obesity[22]. There are now 11 gene regions in which common variation alters type 2 diabetes risk with the levels of statistical confidence that are required by genetic association studies (FIGS 2,3). This progress is all the more remarkable in view of the weak genetic component to type 2 diabetes risk, as compared with many other common diseases that are currently being studied using GWAS. The sibling relative risk is 3–4 at the most for type 2 diabetes, in contrast with 5–10 for rheumatoid arthritis, 15 for type 1 diabetes, 7–10 for bipolar disorder, 17–35 for Crohn disease, 2–7 for early myocardial infarction and 2.5–3.5 for hypertension[23].

The six papers[16–21] describe five separate type 2 diabetes GWAS — extensive replication data from a study in the United Kingdom resulted in a second paper arising from the one initial genome-wide scan. These five studies had several features in common. First, they all used relatively large sample sizes. In combination, DNA samples from more than 18,000 individuals were analysed on the genome-wide chips, with the number of cases ranging from 686 to 1924 and the number of controls from 669 to 5,275 in each study. Second, all studies used DNA samples that were collected from a well defined country or area of Northern Europe, and all participants were of Northern European ancestry. This reduced the likely impact of population admixture — one of the few possible confounding factors that can occur in genetic association studies. Third, all five studies used extensive follow-up case–control studies. The large number of tests a GWAS results in — up to ~400,000 — means that p values of ~5×10^{-7} are needed to provide a study-wide p value of 0.05. The investigators suspected that there would not be many signals that reached this level of significance, so each study assembled between 2,473 and 10,850 additional cases and controls in which to assess their top 'hits'. This brought the total number of cases and controls used in the five studies to approximately 55,000.

Although there were some differences in phenotype definition between studies — some used type 2 diabetic patients of younger age at diagnosis or controls and cases with similar BMI — the overall approaches were remarkably similar. It was therefore reassuring to observe consistent results across the five publications. Five of the six gene regions were reported in at least three papers, and meta-analyses of the individual studies show that the statistical confidence of the findings ranges from 1×10^{-12} to 1×10^{-19} (Supplementary information S1 (box)). The fat mass and obesity-associated (*FTO*) gene region emerged from only one study but, because the association is with BMI, it could not have been detected in studies that used cases and controls of similar BMI. Where more than one publication reported the same locus but a different SNP, the SNPs were always strongly correlated (on the basis of high r^2 values). This showed that the studies had found the same risk allele, even if they had 'tagged' it with a different variant, thereby providing true replication. A summary of the findings is shown in TABLE 1.

What defines a novel gene?

In all cases, the investigators have taken great care to qualify their new findings by saying that they have identified robust 'signals' of association or 'gene regions' rather than actual genes. There are two reasons for this. First, the correlation between common genetic variation (linkage disequilibrium) means that the best association that has been found so far might not represent the causal variant, or combination of variants. For example, in the haematopoietically expressed

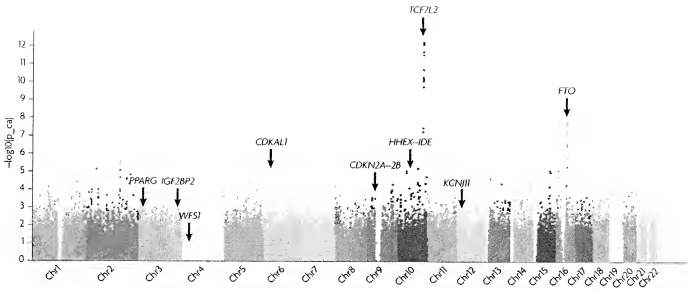


Figure 3 | **Association statistics from one of the five type 2 diabetes genome-wide association studies[20].** The *y* axis represents the −log10 p value and the x axis represents each of the ~400,000 SNPs used in this scan. The point of each arrow indicates the location of the most strongly associated SNP in each of nine known type 2 diabetes gene regions. Two signals, in *SLC30A8* and *TCF2*, were not captured on the Affymetrix chip. The plot was generated using Haploview. *CDKAL1*, CDK5 regulatory-subunit-associated protein 1-like 1; *CDKN2*, cyclin-dependent kinase inhibitor 2A; *FTO*, fat mass and obesity-associated; *HHEX*, haematopoietically expressed homeobox; *IDE*, insulin-degrading enzyme; *IGF2BP2*, insulin-like growth factor 2 mRNA-binding protein 2; *KCNJ11*, potassium inwardly-rectifying channel, subfamily J, member 11; *PPARG*, peroxisome proliferator-activated receptor-γ gene; *SLC30A8*, solute carrier family 30 (zinc transporter), member 8; *TCF2*, transcription factor 2, hepatic; *TCF7L2*, transcription factor 7-like 2 (T-cell specific, HMG-box).

homeobox (*HHEX*)–insulin degrading enzyme (*IDE*) region, strong linkage disequilibrium extends across three genes, two of which are plausible candidates. Second, the functional variant might not lie anywhere near the coding parts of a gene, but might instead influence a regulatory element that might or might not be characterized yet. The *CDKN2A–CDKN2B* locus, which contains cyclin-dependent kinase inhibitor genes, seems to be an example of this: the association is some distance from the nearest genes.

The uncertainty about where the causal variants lie does not detract from the strength of the associations; so, what strength of association convinced the teams that the associations they found were real? Much has been written about what constitutes a real finding, including a recent report from a working group that was set up specifically to address this particular question[23]. One of the main criteria is a p value of approximately $<5\times10^{-7}$. Although this estimate was made on the basis of only a handful of genes so far, this value seems to be about right for type 2 diabetes. For example, in cases in which individual genome scans found evidence of association at $p < 1\times10^{-5}$, the results held up: *TCF7L2* and *FTO* are two examples. By contrast, many signals that reached nominal levels of $p = 1\times10^{-5}$–1×10^{-6} did not hold up when more samples were added, or at least have remained in the 'further studies needed' category. This was the case for the variants in the exostoses (multiple) 2 (*EXT2*) and *LOC387761* genes, which reached p values of 1.8×10^{-5} and 1.2×10^{-5}, respectively, in the genome-scan data reported by Sladek *et al.*, but did not hold up across all the studies. There is further evidence in the form of intermediate trait data — some of the variants are associated with a diabetes-related phenotype in the general population (TABLE 1). This supports the evidence that the associations represent genuine biological findings. Among the new signals identified by GWAS, the most convincing is the association of the *FTO* diabetes risk allele with increased BMI (BOX 1) in the general population, but there is also evidence that the risk alleles in the CDK5 regulatory-subunit-associated protein 1-like 1 gene (*CDKAL1*)[16,19], *HHEX–IDE* (L. Pascoe, E. Ferrannini & M. Walker, personal communication) and the solute-carrier family gene *SLC30A8* (REF 19) regions reduce insulin secretion in healthy adults.

New associations lead to new aetiology

Of the 11 type 2 diabetes gene regions, 4 were found by candidate-gene studies; of the remaining 7, none contains obvious candidate genes. The risk conferred by the individual gene variants is small, but this only reduces the potential predictive value of the risk alleles. A small impact on risk does not detract from the most important aspect of the findings — new associations, provided that they are reproducible, bring new knowledge about disease aetiology. The results of the type 2 diabetes GWAS implicate several pathways involved in β-cell development and function.

Common variants in *TCF7L2* emerged as one of the top signals, if not the top signal, in each study. This is reflected in the estimates of the odds ratios that are conferred by each additional risk allele carried (TABLE 1) at each locus. Each T allele of the key *TCF7L2* SNP, rs7903146, increases disease risk with an odds ratio of 1.37. This is substantially higher than all of the other 10 gene variants, for which the odds ratios range from 1.10 to 1.20. The claims that *TCF7L2* represents, by some distance, the most important type 2 diabetes gene are also reflected by the sample sizes needed to have adequate power to detect the effect. Approximately 1,380 cases and 1,380 controls are needed to detect the *TCF7L2* effect with 80% power at $p = 5\times10^{-7}$, on the basis of allele frequencies in the UK population. The gene region with the next strongest effect, the *CDKN2A–2B* signal, requires 6,200 cases and 6,200 controls. It is still possible that there is a larger signal or a similarly sized signal that has so far not

Table 1 | **Details of 11 type 2 diabetes gene regions**

Example variant	Closest gene	Mode of identification	Previous evidence	Current evidence (p value)*	Additional evidence from human physiology	Odds ratio (per allele)*	RAF (UK)	N‡	Rank in UK GWAS§
rs1801282 (P12A)	*PPARG*	Candidate	Monogenic + drug target	2×10^{-6}	Nothing consistent	1.14 (1.08–1.20)	0.87	>20,000	786
rs5215 (E23K)	*KCNJ11*	Candidate	Monogenic + drug target	5×10^{-13}	Alters insulin secretion in general population	1.14 (1.10–1.19)	0.35	15,600	799
rs7901695	*TCF7L2*	Region-wide	None	1×10^{-55}	Alters insulin secretion in general population	1.37 (1.31–1.43)	0.31	2,760	2
rs4430796	*TCF2*	Candidate	Monogenic	8×10^{-10}	Nothing consistent	1.10 (1.07–1.14)	0.47	>20,000	N/C
rs10010131	*WFS1*	Candidate	Monogenic	1×10^{-7}	Nothing consistent	1.11 (1.08–1.16)	0.60	>20,000	26017
rs1111875	*HHEX–IDE*	Genome-wide	Some, e.g. *HHEX* KO mouse has disrupted pancreatic development	7×10^{-17}	Early studies indicate altered insulin secretion in general population	1.15 (1.10–1.19)	0.65	12,800	32
rs13266634	*SLC30A8*	Genome-wide	None	1×10^{-19}	Early studies indicate altered insulin secretion in general population	1.15 (1.12–1.19)	0.69	14,400	N/C
rs10946398	*CDKAL1*	Genome-wide	None	2×10^{-18}	Early studies indicate altered insulin secretion in general population	1.14 (1.11–1.17)	0.32	16,200	65
rs10811661	*CDKN2A–2B*	Genome-wide	Some – *CDKN2A* KO mouse has reduced islet proliferation	8×10^{-15}	Nothing consistent	1.20 (1.14–1.25)	0.83	12,400	272
rs4402960	*IGF2BP2*	Genome-wide	Some — binds insulin-like growth factor mRNA	9×10^{-16}	Nothing consistent	1.14 (1.11–1.18)	0.32	16,200	1,026
rs8050136	*FTO*	Genome-wide	None	1×10^{-12}	Alters BMI in general population	1.17 (1.12–1.22)	0.40	10,400	12

*Approximate p values and odds ratios calculated by meta-analysis of individual study odds ratios from REFS 16,17,20, except for the signals for *HHEX–IDE*, which also includes data from REF. 18, *CDKAL1*, which also includes data from REF. 19, *SLC30A8*, which includes data from all five GWAS studies, *TCF2*, which is based on data from REFS 8,9 and *WFS1*, which is based on data from REF. 10. ‡Total number of cases and controls needed in a 1:1 ratio to provide 80% power to detect an effect at $p = 5\times10^{-7}$, on the basis of UK risk allele frequencies and assuming a 5% disease frequency. §Position of SNP in the UK genome-wide scan[20] in the list of 393,453 passing quality control and with minor allele frequencies >1%. BMI, body mass index; *CDKAL1*, CDK5 regulatory subunit associated protein 1-like 1; *CDKN2*, cyclin-dependent kinase inhibitor 2A; *FTO*, fat mass and obesity-associated; GWAS, genome-wide association study; *HHEX*, haematopoietically expressed homeobox; *IDE*, insulin-degrading enzyme; *IGF2BP2*, insulin-like growth factor 2 mRNA-binding protein 2; *KCNJ11*, potassium inwardly-rectifying channel, subfamily J, member 11; KO, knockout; N/C, not captured; *PPARG*, peroxisome proliferator-activated receptor-γ gene; RAF, risk allele frequency; *SLC30A8*, solute carrier family 30 (zinc transporter), member 8; *TCF2*, transcription factor 2, hepatic; LF-B3, variant hepatic nuclear factor; *TCF7L2*, transcription factor 7-like 2 (T-cell specific, HMG-box); WFS1, Wolfram syndrome 1.

been identified, because the chips that were used do not cover all of the common and little of the rare variation; however, it seems likely that common variation in *TCF7L2* will represent the most important type 2 diabetes locus in terms of its risk effect and the frequency of the risk allele. Little is known about how *TCF7L2* predisposes to type 2 diabetes. It encodes a transcription factor that is expressed in the fetal pancreas and is involved in the WNT signalling pathway. One of its targets is *HHEX*, which lies in one of the other six identified novel diabetes gene regions. The signal within which *HHEX* falls spans 295 kb and includes three genes: kinesin-interacting factor (*KIF11*) lies in between *HHEX* and *IDE*. *IHHEX*, which encodes a transcription factor with a key role in pancreatic development (knockout mice lack a ventral pancreas[24]), seems to be the most likely candidate out of these three. The association of the diabetes risk allele in the *HHEX–KIF11–IDE* locus with reduced insulin secretion further strengthens the claims that the risk operates through altered β-cell function (Pascoe, Ferrannini & Walker, personal communication).

The association signal on chromosome 9 lies ~120 kb from the 3′ end of *CDKN2B*, which lies next to its close relative *CDKN2A*. Interestingly, recent GWAS have identified a separate set of SNPs that seem to represent the strongest common genetic risk factor for heart disease (myocardial infarction)[21,25,26]. There is no correlation between the diabetes signal and the heart disease signal, but the latter does fall closer to the *CDKN2* genes. *CDKN2A*, which encodes p16INK^{4a}, over-expression of which leads to decreased islet proliferation in ageing mice[27], is the most likely candidate for type 2 diabetes. Initial human physiology studies have not provided any evidence that the risk alleles alter insulin secretion, but the mouse phenotype strongly implicates β-cell dysfunction.

The association of SNPs in *CDKAL1* ranked near the top in four of the five scans. Little is known about *CDKAL1*, but it is highly expressed in human islets[20]. *CDKAL1* shares homology with the CDK5 regulatory-subunit-associated protein-1 gene (*CDK5RAP1*), a known inhibitor of CDK5 activation. CDK5 is implicated in reduced β-cell function, through the

formation of p35–CDK5 complexes, which downregulate insulin expression[28,29]. The association between the diabetes risk allele in the *CDKAL1* locus and reduced insulin secretion further strengthens the claims that altered β-cell function underlies this risk[19].

Less is known about the other genes in the associated regions. Insulin-like growth factor 2 mRNA-binding protein 2 (IGF2BP2) binds to the key growth and insulin signalling molecule insulin-like growth factor 2 (IGF1I) and is also expressed in the pancreatic islet[30]. SLC30A8 is a zinc transporter that is expressed in the β-cell. Of the six new gene regions, the variant in *SLC30A8* that is associated with diabetes risk is the only one that has an obvious functional implication — it is non-synonymous, changing an arginine to a tryptophan.

FTO is the most mysterious of all. A multigene deletion in a mouse model that resulted in a fused-toe phenotype gave the gene its original name[36], but to date the best clue to its role in obesity in humans is its expression in the hypothalamus[22], the key part of the brain that influences appetite.

Finally, deCODE genetics recently showed that the common type 2 diabetes risk alleles in *TCF2* also protect men from prostate cancer ($p = 1 \times 10^{-11}$) (REF. 8). This raises the intriguing possibility that different alleles at the same locus could predispose to cell overgrowth on the one hand and, perhaps, cell degeneration or reduced cell turnover, for example, of the β-cell, on the other hand.

What next?

The five type 2 diabetes GWAS make it clear that human genetic studies are entering a new era. This is almost certainly the case for most common diseases for which results from several GWAS are, or soon will be, in the public domain. Researchers have gone from being able to analyse, at best, a few thousand polymorphisms in a few hundred genes, to more than 80% of the common SNPs in the genome, making it difficult to justify case–control studies that do not have a genome-wide scope. Researchers will soon be able to simply look up the case–control results on the web.

An important next step will be for the geneticists to pass the baton to their clinical, cellular, animal and molecular biology colleagues so that they can work out the mechanisms behind the associations between the new gene variants and disease. But does this mean that human geneticists will be out of the competition? Far from it, there are many areas that need to be addressed. Below are some examples, most of which will apply equally to other diseases.

First, it will be important to fine-map the new type 2 diabetes gene regions. One of the first steps will involve deep sequencing and further rounds of genotyping to build up a full picture of all the possible common variation that might explain the association signals. This should include efforts to define copy number variants such as duplications and deletions, and should also attempt to define independent associations in the same gene regions. The confirmation that a region is involved in disease makes it much more likely that additional variation in the region will predispose to disease. The extent to which African populations, which show reduced linkage disequilibrium, will help to fine-map these regions remains to be seen.

Second, further association studies of the new variants are needed. Investigators will need to assess their role in other populations, especially populations with a high prevalence of diabetes, such as South-Asian, African-American and Mexican-American populations. Further studies of the role of risk alleles in the general population are also important. Type 2 diabetes is associated with many traits, including reduced insulin secretion, insulin resistance, birth weight and

Box 1 | *FTO* gene variants alter fat mass in the general population

Changing lifestyles over the past few decades have caused a rapid increase in the mean body mass index (BMI) in most developed countries. This has had an enormous impact on obesity-related illnesses, including cardiovascular disease, hypertension and possibly even cancer and depression, in addition to type 2 diabetes.

One of the most exciting findings to have come from the genome-wide association studies (GWAS) therefore was that fat mass and obesity associated (*FTO*) genotypes predisposed to type 2 diabetes by altering BMI. The importance of this was highlighted by a separate article that described an association of *FTO* with BMI and obesity risk in the general population[22]. Using ~30,000 adults, the investigators found that the 16% of the European population carrying two copies of the diabetes risk allele were ~1.0 kgm^2 or 2.3 kg heavier than the 35% carrying two copies of the non-risk allele. The statistical confidence behind these findings ($p = 5 \times 10^{-15}$) and its replication in a further study of approximately 8,000 people ($p = 2 \times 10^{-46}$) (REF. 35) make this by far the most convincing evidence for a common gene variant that alters BMI. Even more striking was the finding, using data from more than 5,000 children aged 9 years, that the association was with fat mass, with little or no effect on lean mass[22].

The public health message has been that everybody should eat less and exercise more to reduce their fat mass. The above finding has not changed this message, but does emphasize that some people will find it harder to attain an optimal weight in today's environment than others. Understanding how *FTO* alters fat mass is likely to increase greatly our understanding of obesity.

Glossary

Body mass index
(BMI) BMI is an easy way of estimating how fat people are. It is measured as weight in kilograms divided by height in metres squared (kg/m^2). A BMI < 25 is considered normal, >25 to < 30 overweight and ≥30 obese.

Deep sequencing
Systematic sequencing of contiguous sequence in many individuals.

Linkage disequilibrium
A measure of associations between alleles at different loci, which indicates whether particular haplotypes are more common than expected. We use the r^2 definition which equates directly to power. For example, an r^2 of 0.8 equates to 80% power.

Odds ratio
A measurement of association in case–control studies, defined as the odds of exposure to the susceptibility allele in cases compared with that in controls. If the odds ratio is significantly greater than one, then the allele is associated with an increased risk of the disease.

Population admixture
A process that leads to a composite gene pool in which at least some individuals come from more than one population.

r^2 value
A standard way of quantifying the degree of correlation between polymorphisms. An r^2 of 1 indicates that the two variants would give exactly the same genotypes for an individual.

Sibling relative risk
The chance of being affected by a condition if a sibling is affected, relative to a member of the general population. Siblings of people with type 2 diabetes are three to four times more likely to develop the illness than others.

Tagging
Identifying subsets of markers ('tags') that describe patterns of association or haplotypes among larger marker sets. Tag SNPs are single nucleotide polymorphisms that are correlated with, and therefore can serve as a proxy for, common variation in a region that has not been directly analysed.

inflammation. Preliminary data show that some of the type 2 diabetes risk alleles are associated with reduced insulin secretion in healthy individuals, but more human physiological studies of this type are needed to assess the role of the variants in other traits.

Third, geneticists need to assess the predictive value of the disease variants. Given the small effect sizes that have been seen so far, individual variants will not be useful for predicting type 2 diabetes. However, more extensive studies are needed to assess the usefulness of combining information from multiple variants. The best way of addressing this question is with prospective studies.

Fourth, the potential pharmacogenetic role of new variants needs to be assessed. For many years, the holy grail of common disease genetics has been personalized treatment. Recent data suggest that *TCF7L2* risk alleles predict treatment response to one type of anti-diabetic treatment (sulphonylureas)[31], representing a genuine example of pharmacogenetics, although it is too early to say whether this will be clinically relevant.

Finally, there are more genes to be found. The current findings explain only a small proportion of the excess familial risk. The positions of the four variants found from candidate-gene studies in genome-wide data tell a tantalizing story. Common variants underlying the *PPARG*, *KCNJ11* and *WFS1* associations occurred at positions 786, 799 and 26,017, respectively, in the UK-based genome-wide scan of 1,924 cases and 2,938 controls, and the *TCF2* risk variants were not captured by the Affymetrix chip (TABLE 1). Because each of the GWAS followed up on only a few of the top signals, there must be many more signals to uncover. Extensive collaboration between groups to assemble sample sizes of an order of magnitude of tens of thousands of cases and controls will be necessary to hunt down risk alleles with odds ratios of ~1.10 or less, but which are of no less importance from an aetiological point of view. This is possible. Three of the five type 2 diabetes studies set a benchmark by sharing their summary statistics before publication[16,17,20]. These studies identified the three novel regions found by the other two studies (*SLC30A8*, *HHEX/IDE* and *CDKAL1*) but the *IGF2BP2* and *CDKN2A/2B* regions were revealed only through combining genome-wide data from the total of ~4,500 cases and ~5,500 controls. The study that went on to describe the effect of the *FTO* gene in the general population used a total of ~39,000 samples; a recent study of breast cancer used a total of ~53,000 cases and controls to identify several new risk variants[37].

Conclusion

The first wave of type 2 diabetes GWAS has provided convincing evidence for six novel gene regions involved in the condition. In many cases, the genes in these regions were not previously implicated in the disease. By demonstrating their association, the studies uncovered new aetiological pathways to be explored. This should ultimately lead to improved prevention and treatment for patients. The total number of type 2 diabetes-associated gene regions is now 11, but many more, probably of small effect, remain to be found. Now that sample sizes are in place and technology is not prohibitively expensive, the best way of finding additional genes will be through collaboration to assemble sample sizes of >10,000 cases and controls.

Genetics of Complex Traits, Institute of Biomedical and Clinical Sciences, Peninsula Medical School, University of Exeter, Exeter EX1 2LY, UK.
e-mail: tim.frayling@pms.ac.uk
doi:10.1038/nrg2178

1. Neel, J. V. in *The Genetics of Diabetes Mellitus* (eds Creutzfeldt, W., Kobberling, J. & Neel, J. V.) 1–11 (Springer, Berlin; New York, 1976).
2. O'Rahilly, S., Wainscoat, J. S. & Turner, R. C. Type 2 (non-insulin-dependent) diabetes. New genetics for old nightmares. *Diabetologia* **31**, 407–414 (1988).
3. Horikawa, Y. *et al.* Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. *Nature Genet.* **26**, 163–175 (2000).
4. Florez, J. C. *et al.* Haplotype structure and genotype–phenotype correlations of the sulfonylurea receptor and the islet ATP-sensitive potassium channel gene region. *Diabetes* **53**, 1360–1368 (2004).
5. Gloyn, A. L. *et al.* Large-scale association studies of variants in genes encoding the pancreatic β-cell KATP channel subunits Kir6.2 (*KCNJ11*) and SUR1 (*ABCC8*) confirm that the *KCNJ11* E23K variant is associated with type 2 diabetes. *Diabetes* **52**, 568–572 (2003).
6. Nielsen, E. M. *et al.* The E23K variant of Kir6.2 associates with impaired post-OGTT serum insulin response and increased risk of type 2 diabetes. *Diabetes* **52**, 573–577 (2003).
7. Altshuler, D. *et al.* The common PPARγ Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nature Genet.* **26**, 76–80 (2000).
8. Gudmundsson, J. *et al.* Two variants on chromosome 17 confer prostate cancer risk, and the one in *TCF2* protects against type 2 diabetes. *Nature Genet.* **39**, 977–985 (2007).
9. Winckler, W. *et al.* Evaluation of common variants in the six known maturity-onset diabetes of the young (MODY) genes for association with type 2 diabetes. *Diabetes* **56**, 685–693 (2007).
10. Sandhu, M. S. *et al.* Common variants in *WFS1* confer risk of type 2 diabetes. *Nature Genet.* **39**, 951–953 (2007).
11. Barroso, I. *et al.* Dominant negative mutations in human PPARγ are associated with severe insulin resistance, diabetes mellitus and hypertension. *Nature* **402**, 880–883 (1999).
12. Gloyn, A. L. *et al.* Activating mutations in the gene encoding the ATP-sensitive potassium-channel subunit Kir6.2 and permanent neonatal diabetes. *N. Engl. J. Med.* **350**, 1838–1849 (2004).
13. Inoue, H. *et al.* A gene encoding a transmembrane protein is mutated in patients with diabetes mellitus and optic atrophy (Wolfram syndrome). *Nature Genet.* **20**, 143–148 (1998).
14. Nishigori, H. *et al.* Frameshift mutation, A263fsinsGG, in the hepatocyte nuclear factor-1β gene associated with diabetes and renal dysfunction. *Diabetes* **47**, 1354–1355 (1998).
15. Grant, S. F. *et al.* Variant of transcription factor 7-like 2 (*TCF7L2*) gene confers risk of type 2 diabetes. *Nature Genet.* (2006)
16. Saxena, R. *et al.* Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316**, 1331–1336 (2007).
17. Scott, L. J. *et al.* A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* **316**, 1341–1345 (2007).
18. Sladek, R. *et al.* A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881–885 (2007).
19. Steinthorsdottir, V. *et al.* A variant in *CDKAL1* influences insulin response and risk of type 2 diabetes. *Nature Genet.* **39**, 770–775 (2007).
20. Zeggini, E. *et al.* Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* **316**, 1336–1341 (2007).
21. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
22. Frayling, T. M. *et al.* A common variant in the *FTO* gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* **316**, 889–894 (2007).
23. Chanock, S. J. *et al.* Replicating genotype-phenotype associations. *Nature* **447**, 655–660 (2007).
24. Bort, R., Martinez-Barbera, J. P., Beddington, R. S. & Zaret, K. S. HEX homeobox gene-dependent tissue positioning is required for organogenesis of the ventral pancreas. *Development* **131**, 797–806 (2004).
25. Helgadottir, A. *et al.* A common variant on chromosome 9p21 affects the risk of myocardial infarction. *Science* **316**, 1491–1493 (2007).
26. McPherson, R. *et al.* A common allele on chromosome 9 associated with coronary heart disease. *Science* **316**, 1488–1491 (2007).
27. Krishnamurthy, J. *et al.* p16INK4a induces an age-dependent decline in islet regenerative potential. *Nature* **443**, 453–457 (2006).
28. Ubeda, M., Rukstalis, J. M. & Habener, J. F. Inhibition of cyclin-dependent kinase 5 activity protects pancreatic β cells from glucotoxicity. *J. Biol. Chem.* **281**, 28858–28864 (2006).
29. Wei, F. Y. *et al.* Cdk5-dependent regulation of glucose-stimulated insulin secretion. *Nature Med.* **11**, 1104–1108 (2005).
30. van der Hoeven, F. *et al.* Programmed cell death is affected in the novel mouse mutant *Fused toes* (*Ft*). *Development* **120**, 2601–2607 (1994).
31. Pearson, E. *et al.* Variation in *TCF7L2* influences glycemic response to sulfonylureas. *Diabetes* **56**, 2178–2182 (2007).
32. Easton, D. F. *et al.* Genome-wide association study identifies novel breast cancer susceptibility loci *Nature* **447**, 1087–1093 (2007).
33. Dina, C. *et al.* Variation in *FTO* contributes to childhood obesity and severe adult obesity. *Nature Genet.* **39**, 724–726 (2007).

Acknowledgements
I would like to thank specifically M. Weedon and R. Freathy for help and advice on the article, and more generally all my colleagues in Exeter and Oxford with whom I worked on the UK type 2 diabetes genome-wide study.

Competing interests statement
The authors declare no competing financial interests.

DATABASES
Entrez Gene: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene
CDKAL1 | CDKAL1 | CDKN2A | CDKN2B | FTO | HHEX | IDE | IGF1 | KCNJ11 | PPARG | SLC30A8 | TCF7L2 | WFS1
OMIM: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM
bipolar disorder | Crohn disease | hypertension | early-onset atrial fibrillation | Hutchinson–Gilford progeria syndrome | type 2 diabetes | Wolfram syndrome
UniProtKB: http://ca.expasy.org/sprot
IGF2BP2

FURTHER INFORMATION
deCODE genetics: http://www.decode.com
Haploview: http://www.broad.mit.edu/mpg/haploview/

SUPPLEMENTARY INFORMATION
See online article: S1 (box)

ALL LINKS ARE ACTIVE IN THE ONLINE PDF.

Zeggini et al (Nature Genetics Mar 30 [Epub ahead of print] (2008))

nature genetics

Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes

Eleftheria Zeggini[1,10], Laura J Scott[2,10], Richa Saxena[3–8,10] & Benjamin F Voight[3–5,7,10], for the Diabetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium[9]

Genome-wide association (GWA) studies have identified multiple loci at which common variants modestly but reproducibly influence risk of type 2 diabetes (T2D)[1–11]. Established associations to common and rare variants explain only a small proportion of the heritability of T2D. As previously published analyses had limited power to identify variants with modest effects, we carried out meta-analysis of three T2D GWA scans comprising 10,128 individuals of European descent and ~2.2 million SNPs (directly genotyped and imputed), followed by replication testing in an independent sample with an effective sample size of up to 53,975. We detected at least six previously unknown loci with robust evidence for association, including the *JAZF1* ($P = 5.0 \times 10^{-14}$), *CDC123-CAMK1D* ($P = 1.2 \times 10^{-10}$), *TSPAN8-LGR5* ($P = 1.1 \times 10^{-9}$), *THADA* ($P = 1.1 \times 10^{-9}$), *ADAMTS9* ($P = 1.2 \times 10^{-8}$) and *NOTCH2* ($P = 4.1 \times 10^{-8}$) gene regions. Our results illustrate the value of large discovery and follow-up samples for gaining further insights into the inherited basis of T2D.

GWA studies are unbiased by previous hypotheses concerning candidate genes and pathways, but they are limited by the modest effect sizes of individual common susceptibility variants and the need for stringent statistical thresholds. For example, the largest allelic odds ratio (OR) of any established common variant for T2D is ~1.35 (*TCF7L2*), and the nine other validated associations to common variants (excluding *FTO*, which has its primary effect through obesity) have allelic ORs between 1.1 and 1.2 (refs. 1–6,11,12). To augment power to detect additional loci of similar or smaller effect, we increased sample size by combining three previously published GWA studies (Diabetes Genetics Initiative (DGI), Finland–United States Investigation of NIDDM Genetics (FUSION) and Wellcome Trust Case Control Consortium (WTCCC))[1–4], and extended SNP coverage by imputing untyped SNPs on the basis of patterns of haplotype variation from HapMap[13] (**Table 1**).

We started with a set of genotyped autosomal SNPs that passed quality control filters in each study: in WTCCC, 393,143 SNPs from the Affymetrix 500K chip (minor allele frequency (MAF) > 0.01; 1,924 cases and 2,938 population-based controls[3,4]); in DGI, 378,860 SNPs from the Affymetrix 500K chip (MAF > 0.01; Swedish and Finnish sample of 1,464 T2D cases and 1,467 normoglycemic controls, including 326 discordant sibships[1]); and in FUSION, 306,222 SNPs from the Illumina 317K chip (MAF > 0.01, 1,161 T2D cases and 1,174 normal glucose-tolerant controls from Finland[2]) (**Supplementary Table 1** online). 44,750 SNPs (MAF > 0.01) were directly genotyped in all three studies across the two platforms. We used data from the GWA studies and phased chromosomes from the HapMap CEU sample to impute autosomal SNPs with MAF > 0.01 (ref. 14; see also URLs section in Methods). We based our further analyses on 2,202,892 SNPs that met imputation and genotyping quality control criteria across all studies (**Supplementary Methods** online).

Using these directly measured and imputed genotypes, we tested for association of each SNP with T2D in each study separately, corrected each study for residual population stratification, cryptic relatedness or technical artifacts using genomic control, and then combined these results in a genome-wide meta-analysis across a total of 10,128 samples (4,549 cases and 5,579 controls; **Supplementary Methods**). We calculated that this sample size provides reasonable power to detect additional variants with properties similar to those previously identified through less formal data combination efforts[1,2,4] (**Supplementary Table 2** online). Unless otherwise indicated, results presented are derived from individually genomic control–adjusted stage 1 results. We obtained meta-analysis OR and confidence intervals

[1]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. [2]Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan 48109, USA. [3]Broad Institute of Harvard and Massachusetts Institute of Technology (MIT), Cambridge, Massachusetts 02142, USA. [4]Center for Human Genetic Research, [5]Department of Medicine and [6]Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [7]Department of Medicine and [8]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. [9]The full list of authors and affiliations appears at the end of this paper. [10]These authors contributed equally to this work. Correspondence should be addressed to M.I.M. (mark.mccarthy@drl.ox.ac.uk), M.B. (boehnke@umich.edu) or D.A. (altshuler@molbio.mgh.harvard.edu).

Received 18 December 2007; accepted 12 February 2008; published online 30 March 2008; doi:10.1038/ng.120

Table 1 Overview of study design

Study	Cases (*n*)[a]	Controls (*n*)[a]	Effective sample size[a]	Number of directly genotyped SNPs[b]	Number of imputed SNPs[b]
Stage 1					
DGI	1,464	1,467	2,521	378,860	1,888,145
WTCCC	1,924	2,938	4,706	393,143	1,915,393
FUSION	1,161	1,174	2,335	306,222	2,110,199
Stage 2					
DGI stage 2	5,065	5,785	9,874	63	–
FUSION stage 2	1,215	1,258	2,473	59	–
UK stage 2	3,757	5,346	9,114	66	–
Stage 3					
deCODE	1,520 (1,422)	25,235 (3,455)	4,280 (3,130)	11	–
KORA	1,241	1,458	2,684	6	–
Danish	4,089	5,043	8,690	11	–
HUNT	1,004	1,503	2,412	11	–
NHS	1,506	2,014	3,468	10	–
CCC	547	533	1,070	11	–
EPIC	388	774	1,036	10	–
ADDITION/Ely	892	1,610	2,288	11	–
Norfolk	2,311	2,400	4,450	11	–
METSIM	659	2,639	2,136	11	–

[a]Sample sizes presented here are the maximum available for each study. For the deCODE stage 3 study, we used genotype data from the Icelandic GWA scan[5] for rs2641348, rs7578597 and rs9472138, and a perfect proxy (rs2793831, based on HapMap) for rs10923931. The remaining SNPs had not been directly typed as part of this scan and were therefore genotyped separately, in a subset of the GWA scan samples (numbers indicated in parentheses) (**Supplementary Methods**). [b]Autosomal SNPs passing quality control, as defined for directly genotyped and imputed SNPs in each study (quality control criteria: SNPTEST information measure >0.5; r^2hat ≥ 0.3; MAF > 0.01). For the stage 1 meta-analyses, we combined results for 2,202,892 directly genotyped and imputed SNPs passing quality control in all three studies (**Supplementary Methods**).

from a fixed-effects model, and *P* values from a weighted *z* statistic–based meta-analysis (**Supplementary Methods**). As expected, the most significant result was obtained for rs7903146 in *TCF7L2*. We also observed evidence for association ($P < 10^{-3}$) at eight of the ten established T2D loci (as well as at the *FTO* obesity locus)[12] (**Supplementary Table 3** online). This was unsurprising, as these same data supported the identification of many of these loci. As our goal was to identify previously unknown loci, we excluded 1,981 SNPs in the immediate vicinity of these T2D susceptibility loci from further analysis (with the exception of a signal near *PPARG*, which was followed up), and examined the remainder of the autosomal genome (**Supplementary Methods**). Even after excluding known loci, we saw a strong enrichment of highly associated variants: 426 with *P* values $<10^{-4}$, compared to 217 under the null.

Before proceeding to follow-up, we explored the individual studies and the combined data for potential errors and biases. We found a genomic control λ value of 1.04 for the combined results (based on 10,128 samples), which, given the relationship between λ and sample size[15], suggests little residual confounding (**Supplementary Fig. 1** and **Supplementary Note** online). We also used genome-wide genotype data to estimate the principal components of the identity-by-state relationships in each stage 1 sample. For the SNPs presented in **Table 2**, adjustment for principal components in stage 1 T2D association analysis did not diminish the association in the WTCCC (two principal components), FUSION (ten principal components) or DGI (ten principal components) samples (**Supplementary Note**). Additionally, we did not find any evidence for association between UK population ancestry informative markers[3] and disease status in the UK replication sets (**Supplementary Note**). To ensure that the observed stage 1 associations taken forward to follow-up were not due to imputation errors, we directly genotyped originally imputed variants in the stage 1 samples (**Supplementary Methods**). We found strong agreement between the genotype-based and imputed *P* values: in 38 of 43 cases where a direct genotype-based result was obtained, the *P* value was within one order of magnitude of that derived from imputation, and in the remaining five cases, *P* values were less than two orders of magnitude different (**Supplementary Table 4** online).

We selected SNPs for replication principally on the basis of the statistical evidence for association in stage 1, excluding SNPs with evidence for heterogeneity of ORs ($P < 10^{-4}$) across studies (**Supplementary Methods**). We took 69 SNPs forward to an initial round of replication (stage 2) in up to 22,426 additional samples of European descent (**Table 1** and **Supplementary Table 1**). The distribution of association *P* values in stage 2 was highly inconsistent with a null distribution. Of the 69 signals selected for follow up, 65 were successfully genotyped in stage 2, and represented loci that were independent of each other and of previously established susceptibility loci. Nine of these had a *P* value ≤0.01 with association in the same direction as the original signal, far in excess of the 0.33 expected under the null ($P = 1.4 \times 10^{-12}$, binomial test; **Supplementary Methods**), and two SNPs had $P < 10^{-4}$ as compared to an expectation of 0.0033 ($P = 5.2 \times 10^{-6}$) (**Supplementary Methods** and **Supplementary Table 5** online).

We identified 11 SNPs (ten separate signals, nine of which represent previously unknown loci) with $P < 0.005$ in stage 2 for which the combined stage 1 and stage 2 data (based on direct genotyping of stage

Table 2 Eleven T2D-associated SNPs taken forward to stages 2 and 3

							Stage 1 (DGI, FUSION, WTCCC)		Stage 2 (DGI, FUSION, UKT2D)		Stage 3 (deCODE, KORA, Steno, HUNT, NHS, CCC, EPIC, ADDITIONEly, Norfolk, METSIM)		All data				
SNP	Chr	Position NCBI35 (bp)	Nonrisk allele[a]	Risk allele[a]	Risk allele frequency[a]	Nearest gene(s)	OR (95% CI)	*P* value	OR (95% CI)	*P* value	OR (95% CI)	*P* value	n_{eff}	OR (95% CI)	*P* value	P_{het}	Number of samples for 80% power[b]
rs864745	7	27,953,796	C	**T**	0.501	*JAZF1*	1.14 (1.07–1.20)	1.5E-04	1.08 (1.04–1.12)	8.1E-05	1.10 (1.06–1.15)	1.3E-07	59,617	1.10 (1.07–1.13)	5.0E-14	0.70	10,610
rs12779790	10	12,368,016	**A**	G	0.183	*CDC123*, *CAMK1D*	1.15 (1.06–1.24)	4.2E-04	1.11 (1.06–1.16)	5.4E-05	1.09 (1.04–1.14)	1.5E-04	62,366	1.11 (1.07–1.14)	1.2E-10	0.67	9,334
rs7961581	12	69,949,369	**T**	C	0.269	*TSPAN8*, *LGR5*	1.18 (1.10–1.26)	1.8E-05	1.06 (1.02–1.11)	9.8E-03	1.09 (1.04–1.13)	4.3E-05	62,301	1.09 (1.06–1.12)	1.1E-09	0.20	23,206
rs7578597	2	43,644,474	C	**T**	0.902	*THADA*	1.25 (1.12–1.40)	1.8E-04	1.15 (1.07–1.22)	1.6E-03	1.12 (1.05–1.20)	9.2E-05	60,832	1.15 (1.10–1.20)	1.1E-09	0.008	9,624
rs4607103	3	64,686,944	T	**C**	0.761	*ADAMTS9*	1.13 (1.06–1.22)	5.4E-04	1.10 (1.05–1.15)	1.0E-04	1.06 (1.01–1.11)	3.5E-03	62,387	1.09 (1.06–1.12)	1.2E-08	0.17	9,748
rs10923931[c]	1	120,230,001	**G**	**T**	0.106	*NOTCH2*	1.30 (1.17–1.43)	1.1E-04	1.09 (1.03–1.16)	2.9E-03	1.11 (1.05–1.18)	1.9E-03	58,667	1.13 (1.08–1.17)	4.1E-08	0.004	21,568
rs1153188	12	53,385,263	T	**A**	0.733	*DCD*	1.15 (1.08–1.23)	3.2E-05	1.07 (1.03–1.12)	3.1E-03	1.06 (1.02–1.10)	8.8E-03	62,301	1.08 (1.05–1.11)	1.8E-07	0.79	17,808
rs17036101[d]	3	12,252,845	**A**	G	0.927	*SYN2*, PPARG	1.33 (1.18–1.50)	1.0E-05	1.13 (1.04–1.22)	4.5E-03	1.11 (1.02–1.20)	1.2E-02	59,682	1.15 (1.10–1.21)	2.0E-07	0.19	16,370
rs2641348[c]	1	120,149,926	**A**	G	0.107	*ADAM30*	1.14 (1.05–1.25)	1.4E-03	1.10 (1.03–1.17)	1.2E-03	1.09 (1.03–1.16)	7.8E-03	60,048	1.10 (1.06–1.15)	4.0E-07	0.08	17,428
rs9472138	6	43,919,740	**C**	**T**	0.282	*VEGFA*	1.13 (1.06–1.21)	5.4E-05	1.07 (1.02–1.12)	1.5E-03	1.03 (1.00–1.07)	9.5E-02	63,537	1.06 (1.04–1.09)	4.0E-06	0.43	16,696
rs10490072	2	60,581,582	C	**T**	0.724	*BCL11A*	1.17 (1.10–1.26)	3.4E-05	1.08 (1.03–1.13)	1.4E-03	1.00 (0.97–1.04)	6.5E-01	59,682	1.05 (1.03–1.08)	1.0E-04	0.0035	13,502
Maximum available effective sample size							9,562		21,461		32,514						

Results from the analysis of directly genotyped data only, except for FUSION stage 1 results for rs7961581 (**Supplementary Methods**). Combined estimates of ORs were calculated using a fixed effects, inverse variance meta-analysis; DGI discordant sibling pairs were not included in OR estimates. *P* values were combined using a weighted *z* score–based meta-analysis including DGI sibships; *P* values for the three stage 1 studies were individually corrected by genomic control before meta analysis.

[a]Ancestral allele is denoted in bold, based on Entrez SNP and derived by comparison against chimpanzee sequence. The risk allele frequencies presented are sample size-weighted risk allele frequencies across the stage 2 studies. [b]Sample size (sum of case and control samples) required for 80% power (to achieve nominal replication at $\alpha = 0.05$) is calculated on the basis of the stage 2 OR estimate, sample size–weighted risk allele frequency across the stage 2 studies and assuming an equal number of cases and controls (**Supplementary Methods**). [c]SNPs rs10923931 and rs2641348 appear to represent the same signal ($r^2 = 0.92$ in HapMap CEU)[12]. Results for rs2934381 and rs2793831 (perfect proxies for rs10923931) are presented for UK (stage 1,2) and deCODE (stage 3) respectively. [d]The signal at SNP rs17036101 is indistinguishable from that at rs1801282, the established P12A variant in *PPARG*.

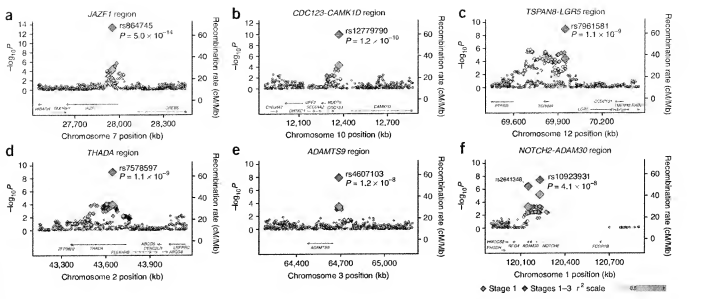


Figure 1 Regional plots of six confirmed associations. (**a**–**f**) For each of the *JAZF1* (**a**), *CDC123-CAMK1D* (**b**), *TSPAN8-LGR5* (**c**), *THADA* (**d**), *ADAMTS9* (**e**) and *NOTCH2-ADAM30* (**f**) regions, genotyped and imputed SNPs passing quality control across all three stage 1 studies are plotted with their meta-analysis *P* values (as $-\log_{10}$ values) as a function of genomic position (NCBI Build 35). In each panel, the SNP taken forward to stages 2 and 3 is represented by a blue diamond (meta-analysis *P* value across stages 1–3), and its initial *P* value in stage 1 data is denoted by a red diamond. Estimated recombination rates (taken from HapMap)[13] are plotted to reflect the local LD structure around the associated SNPs and their correlated proxies (according to a white to red scale from $r^2 = 0$ to $r^2 = 1$; based on pairwise r^2 values from HapMap CEU)[13]. Gene annotations were taken from the University of California Santa Cruz Genome Browser.

I samples, where previously imputed) generated $P < 10^{-5}$. We further genotyped these 11 SNPs in up to 57,366 additional samples (14,157 cases and 43,209 controls) of European descent in stage 3 (**Table 1, Supplementary Table 1** and **Supplementary Methods**). The distribution of *P* values for these 11 SNPs was again inconsistent with a null distribution: all nine newly identified and independent SNPs had effects in the same direction as in the stage 1 + 2 meta-analysis ($P = 0.002$), and seven had $P < 0.05$ in the direction of the original association ($P = 2.1 \times 10^{-10}$) (**Table 2**).

On the basis of the combined stage 1–3 analyses, we found that six signals reached compelling levels of evidence ($P = 5.0 \times 10^{-8}$ or better) for association with T2D (**Table 2**). As in all linkage disequilibrium (LD)-mapping approaches, characterization of the causal variants responsible, their effect sizes and the genes through which they act will require extensive resequencing and fine-mapping. However, on the basis of current evidence, we found that the most associated variants in each of these signals map to intron 1 of *JAZF1*, between *CDC123* and *CAMK1D*, between *TSPAN8* and *LGR5*, in exon 24 of *THADA*, near *ADAMTS9* and in intron 5 of *NOTCH2*.

The strongest statistical evidence for a new association signal was for rs864745 in intron 1 of *JAZF1* (**Fig. 1**), one of a cluster of associated SNPs with strong evidence for association in the stage 1 meta-analysis and across each replication sample (**Table 2** and **Supplementary Table 6** online). The overall estimate of effect was an OR of 1.10 (95% CI = 1.07–1.13; $P = 5.0 \times 10^{-14}$ under an additive model), based on 68,042 individuals. *JAZF1* (juxtaposed with another zinc finger gene 1) encodes a transcriptional repressor of NR2C2 (nuclear receptor subfamily 2, group C, member 2)[16]. Mice deficient in *Nr2c2* show growth retardation, low IGF1 serum concentrations and perinatal and early postnatal hypoglycaemia[17]. Very recently, a SNP in *JAZF1* was identified as associated with prostate cancer[18]; this is particularly interesting given the recent finding that SNPs in *HNF1B* are also associated both with T2D and prostate cancer[19,20].

The second strongest statistical evidence for a new signal was for rs12779790 (combined OR = 1.11, 95% CI = 1.07–1.14, $P = 1.2 \times 10^{-10}$), which lies in an intergenic region ~90 kb from *CDC123* (cell division cycle 123 homolog (*S. cerevisiae*)) and ~63.5 kb from *CAMK1D* (calcium/calmodulin-dependent protein kinase ID) (**Fig. 1, Table 2** and **Supplementary Table 6**). CDC123 is regulated by nutrient availability in *S. cerevisiae* and has a role in cell cycle regulation[21]. Evidence from previous GWA studies implicating variants in *CDKAL1* and near *CDKN2A/B* in T2D predisposition suggests that cell cycle dysregulation may be a common pathogenetic mechanism in T2D[1,2,4].

The third strongest statistical signal was found for rs7961581, which resides upstream of *TSPAN8* (tetraspanin 8; combined OR = 1.09, 95% CI = 1.06–1.12, $P = 1.1 \times 10^{-9}$) (**Fig. 1, Table 2** and **Supplementary Table 6**). Tetraspanin 8 is a cell-surface glycoprotein expressed in carcinomas of the colon, liver and pancreas.

The fourth strongest new association signal was found for rs7578597, a nonsynonymous SNP (T1187A; combined OR = 1.15, 95% CI = 1.10–1.20, $P = 1.1 \times 10^{-9}$) that resides in exon 24 of the widely expressed *THADA* (thyroid adenoma associated) gene (**Fig. 1, Table 2** and **Supplementary Table 6**). Disruption of *THADA* by chromosomal rearrangements (including fusion with intronic sequence from *PPARG*) is observed in thyroid adenomas[22]. The function of THADA has not been well characterized, but there is some evidence to suggest it may be involved in the death receptor pathway and apoptosis[23].

rs4607103 (combined OR = 1.09, 95% CI = 1.06–1.12, $P = 1.2 \times 10^{-8}$), representing a cluster of associated SNPs, resides ~38 kb upstream of *ADAMTS9* (ADAM metallopeptidase with thrombospondin

type 1 motif, 9), and is the SNP with the fifth strongest signal (**Fig. 1**, Table 2 and **Supplementary Table 6**). ADAMTS9 is a secreted metalloprotease that cleaves the proteoglycans versican and aggrecan, and it is expressed widely, including in skeletal muscle and pancreas.

The sixth strongest signal is marked by rs10923931, which resides within intron 5 of *NOTCH2* (Notch homolog 2 (*Drosophila*); combined OR = 1.13, 95% CI = 1.08–1.17, $P = 4.1 \times 10^{-8}$) (**Fig. 1**, Table 2 and **Supplementary Table 6**). We also followed up on rs2641348, a nonsynonymous SNP (L359P) within the neighboring gene *ADAM30* (ADAM metallopeptidase domain 30) that represents the same signal ($r^2 = 0.92$ based on HapMap CEU data), but we found that its overall signal (combined OR = 1.10, 95% CI = 1.06–1.15, $P = 4.0 \times 10^{-7}$; **Table 2**) was slightly weaker. NOTCH2 is a type 1 transmembrane receptor; in mice, Notch2 is expressed in embryonic ductal cells of branching pancreatic buds during pancreatic organogenesis, the likely source of endocrine and exocrine stem cells[24].

The strength of the association evidence for the remaining four variants taken into stage 3 did not meet our prespecified threshold of $P \leq 5.0 \times 10^{-8}$. However, it is likely (based on individual significance values and their overall distribution) that several of these variants also represent genuine association signals. In all, three of these additional SNPs showed P values $< 10^{-5}$ across the combined data (**Table 2**), and two had $P < 0.05$ in stage 3 in the same direction as in stages 1 and 2. Variants near *DCD* (dermcidin) showed evidence for association (rs1153188; overall $P = 1.8 \times 10^{-7}$) (**Supplementary Fig. 2** online). A signal in *VEGFA* had previously been noted in the WTCCC GWA scan[4], but it showed inconsistent evidence for replication: further studies will be required to establish its status. We also found association at rs17036101, ~44 kb downstream of *SYN2* (synapsin II) and 115.3 kb upstream of the established T2D susceptibility variant rs1801282 (P12A) in *PPARG* ($r^2 = 0.54$ in HapMap CEU) (**Supplementary Fig. 3** online). Conditional analyses in stage 1 + 2 samples could not differentiate between the effect of these two SNPs (**Supplementary Note** and **Supplementary Table 7** online).

None of the 11 SNPs (**Table 2**) were convincingly associated with body mass index (BMI) (**Supplementary Table 8** online) or other T2D-related traits (with $P < 10^{-3}$) (**Supplementary Table 9** online). The largest fold-change in T2D association P values before adjusting for BMI was for rs17036101 ($P = 8.1 \times 10^{-8}$ before adjustment and $P = 7.5 \times 10^{-6}$ after adjustment for BMI; **Supplementary Table 10** online). Conditioning on the associated SNP that was taken forward to stages 2 and 3 in each region showed no additional independent association signals ($P < 10^{-4}$) in stage 1 data (**Supplementary Note** and **Supplementary Fig. 4** online).

By combining three GWA scans involving 10,128 samples (enhanced through imputation approaches) and undertaking large-scale replication in up to 79,792 additional samples, we identified six additional loci that apparently harbor common genetic variants influencing susceptibility to T2D. These findings are consistent with a model in which the preponderance of loci detectable through the GWA approach (using current arrays and indirect LD mapping) have modest effects (ORs between 1.1 and 1.2). Given such a model, our study (in which we followed up only 69 signals out of over 2 million meta-analysed SNPs) would be expected to recover only a subset of the loci with similar characteristics (that is, those that managed to reach our stage 1 selection criteria). Further efforts to expand GWA meta-analyses and to extend the number of SNPs taken forward to large-scale replication should confirm additional genomic loci, as should targeted analysis of copy number variation. However, the present data provide only crude estimates of the overall effect on susceptibility attributable to variants at these loci. The effect of the actual common causal variant responsible for the index association (once identified) will typically be larger, and many of these loci are likely to carry additional causal variants, including, on occasion, low-frequency variants of larger effect: three genes with common variants that influence risk of T2D were first identified on the basis of rare mendelian mutations (in *KCNJ11*, *WFS1* and *HNF1B*). Regardless of effect size, these loci provide important clues to the processes involved in the maintenance of normal glucose homeostasis and in the pathogenesis of T2D.

METHODS

Stage 1 samples, genome-wide genotyping and quality control. An expanded description of these methods is provided in **Supplementary Methods**.

The WTCCC stage 1 sample consists of 1,924 T2D cases and 2,938 population controls from the UK[3,4]. These samples were genotyped on the Affymetrix GeneChip Human Mapping 500K Array Set. The call frequency of included samples was >0.97. In total, 393,143 autosomal SNPs passed quality control criteria (Hardy-Weinberg equilibrium (HWE) $P > 10^{-4}$ in T2D cases and controls; call frequency >0.95, MAF > 0.01 and good clustering[3,4]).

The DGI stage 1 Swedish and Finnish sample consists of 1,464 T2D cases and 1,467 normoglycemic controls. Of these, 2,097 are population-based T2D cases and controls matched for body mass index (BMI), gender and geographic origin, and 834 are T2D cases and controls in 326 sibships discordant for T2D[1]. These samples were genotyped on the Affymetrix GeneChip Human Mapping 500K Array Set, and all included samples had a genotype call rate >0.95. In total, 378,860 autosomal SNPs passed quality control criteria (call frequency >0.95, HWE $P > 10^{-6}$ in controls and MAF > 0.01 in both population and familial components)[1].

The FUSION stage 1 sample consists of 1,161 Finnish T2D cases and 1,174 Finnish normal glucose-tolerant controls[2]. In addition, we included 122 FUSION offspring with genotyped parents for quality control purposes and quantitative trait analysis. Samples were genotyped with the Illumina HumanHap300 BeadChip (v1.1). All samples included had a call frequency >0.975. In sum, 306,222 autosomal SNPs passed quality control (HWE $P \geq 10^{-6}$ in the total sample, ≤3 combined duplicate or nonmendelian inheritance errors (out of 79 duplicate samples and 122 parent–offspring sets), call frequency ≥0.90 and MAF > 0.01) (ref. 2).

Analysis of stage 1 genotype data. In combining data across the three studies, we did not attempt, given differences in study design and implementation, to harmonize every aspect of individual study analysis and quality control. For the UK, DGI and FUSION studies, respectively, 393,143, 378,860 and 306,222 SNPs were analyzed under an additive model. The genomic control values for these directly genotyped SNPs were 1.08 (UK), 1.06 (DGI) and 1.03 (FUSION) (**Supplementary Methods**).

Stage 1 imputation and T2D analysis. For each stage 1 sample set, we imputed genotypes for autosomal SNPs that were present in HapMap Phase II but that were not present in the genome-wide chip or that did not pass direct genotyping quality control. In each sample, genotypes were imputed using the genotype data from the GWA chips and phased HapMap II genotype data from the 60 CEU HapMap founders. We retained SNPs that had an estimated MAF > 0.01 in the control or total sample. Imputed SNPs were then tested for T2D association. The genomic control values for these imputed SNPs were 1.08 (UK), 1.07 (DGI) and 1.04 (FUSION) (**Supplementary Methods**).

Stage 1 meta-analysis. An expanded description of these methods is provided in **Supplementary Methods**. We used meta-analysis to combine the T2D association results for the stage 1 WTCCC, DGI and FUSION samples. The combined stage 1 data are comprised of 10,128 samples: 4,549 T2D cases and 5,579 controls. We used association results from directly genotyped SNPs, where available, and imputed genotype association results at all other positions. 2,202,892 genotyped and imputed autosomal SNPs passed quality control and had MAF > 0.01 in each of the three samples (44,750 were genotyped in all three samples, 308,685 were genotyped in two samples, 250,280 were genotyped in one sample, and 1,599,177 were imputed in all samples). All association

results were expressed relative to the forward strand of the reference genome based on dbSNP125. In our initial analysis, which was used to select signals for stage 2 genotyping, for each SNP we combined the ORs for a given reference allele weighted by the confidence intervals using a fixed effects model. We investigated evidence for heterogeneity of ORs using two commonly used statistics: Cochran's Q statistic and I^2 (ref. 25).

We repeated the meta-analysis, combining evidence for association solely on the basis of the P values. Specifically, for each study, we converted the two-sided P value to a z statistic that was signed to reflect the direction of the association given the reference allele. Each z score was then weighted; the squared weights were chosen to sum to 1, and each sample-specific weight was proportional to the square root of the effective number of individuals in the sample. We summed the weighted z statistics across studies and converted the summary z score to a two-sided P value.

SNP prioritization for stage 2 genotyping. We prioritized 69 SNPs for replication in stage 2 on the basis of the results from the three-study stage 1 meta-analysis, using a set of criteria we developed as part of a heuristic approach to the prioritization of loci for follow-up (**Supplementary Methods**). We considered SNPs with a meta-analysis P value $<10^{-4}$ and a meta-analysis heterogeneity P value $>10^{-4}$. These selections were largely made using the initial OR-based version of the meta-analysis. We allowed some exceptions to the above follow-up criteria.

Five SNPs were selected for replication genotyping on the basis of their strong association with T2D in the DGI GWA study (two SNPs), association with T2D and with insulinogenic index in the DGI study (one SNP), and overlap with FUSION or WTCCC ($P < 0.05$ in DGI and one or both studies; two SNPs). For known T2D loci (*TCF7L2*, *CDKAL1*, *IGF2BP2*, *KCNJ11*, *HHEX/IDE*, *SLC30A8*, *CDKN2A/B* region, *WFS1*, *HNF1B* and *FTO*), we excluded from follow-up all SNPs that resided within the surrounding region, with region boundaries defined by the furthest neighboring SNPs with P values remaining ~ 0.01 ($n = 1{,}981$). For the *PPARG* region, we identified a SNP, rs17036101, with a P value two orders of magnitude lower than the established P12A susceptibility variant, rs1801282, and we took this signal forward to replication. In total, we took 69 SNPs forward to stage 2 genotyping.

Stage 2 samples, genotyping and analysis. We genotyped the prioritized SNPs in cases and controls from three UK replication sets (RS1, RS2 and RS3, described in ref. 4; **Supplementary Table 1** and **Supplementary Methods**). Genotyping of prioritized SNPs in RS1, RS2 and RS3 was done by KBiosciences. All assays were validated prior to use, using a standard 96-well validation plate (KBiosciences) and up to 296 samples from the WTCCC study (**Supplementary Methods**). Concordance rates between the Affymetrix and KASPar/TaqMan genotypes (based on up to 296 replicate stage 1 samples) were 97.5% on average. All genotyped SNPs had genotype call frequency rates >94% in the replication sets, and no SNPs had HWE $P < 0.001$ in cases or controls. We tested for association with T2D using the Cochran-Armitage test for trend. Results from the three replication sets were combined in a Cochran-Mantel-Haenszel meta-analysis framework.

For DGI, we genotyped the prioritized SNPs in three stage 2 case-control samples[1] (**Supplementary Table 1** and **Supplementary Methods**). The prioritized SNPs were genotyped in all DGI stage 1 and 2 samples using the iPLEX Sequenom MassARRAY platform. We used 63 SNPs passing quality control (>94% call rate, MAF > 0.01 and HWE P value >0.001) for association testing. We tested for T2D association in each DGI stage 2 case-control set using a χ^2 analysis (assuming an additive genetic model). Results from the three DGI stage 2 samples were combined using Cochran-Mantel-Haenszel meta-analysis.

For FUSION, we genotyped the prioritized SNPs in a Finnish case-control sample (**Supplementary Table 1** and **Supplementary Methods**) using the Sequenom Homogeneous Mass EXTEND or iPLEX Gold SBE assays, carried out at the National Human Genome Research Institute (NHGRI). In sum, 59 SNPs had genotype call frequency >94% and HWE P value >0.001. The genotype consistency rate among 56 duplicate samples was 100%, and the average call frequency of successfully genotyped SNPs was 97.3%. SNPs were analyzed using logistic regression with adjustment for sex, 5-year age category and birth province and an additive model for the genetic effect.

Comparison of genotypes from imputation and direct genotyping. We genotyped a proportion of the prioritized imputed signals in the stage 1 samples of the three studies, and calculated respective concordance rates (**Supplementary Methods** and **Supplementary Table 4**). All results presented in the main manuscript text are based on directly typed stage 1 data, except rs7961581 in FUSION stage 1.

Combined meta-analysis for stages 1 and 2. We combined stage 1 and stage 2 data using both the OR-based and the weighted z score–based meta-analysis approaches described above. We also assessed our results using random effects meta-analysis to better account for any heterogeneity between the studies (**Supplementary Table 6**). Locus-specific and combined sibling relative risk estimates were calculated using sample size–weighted estimates of the effect size and risk-allele frequency derived from stage 2 replication samples only, and under the assumption of allelic and locus independence, as described[26,27].

Stage 3 sample, genotyping and association analysis. We followed up 11 SNPs (rs2641348, rs10490072, rs7578597, rs17036101, rs4607103, rs9472138, rs864745, rs12779790, rs1153188, rs10923931 and rs7961581) in stage 3 samples from the deCODE, KORA, Danish, HUNT, NHS, GEM Consortium (CCC, EPIC, ADDITION/Ely, Norfolk) and METSIM studies (**Supplementary Table 1** and **Supplementary Methods**).

Combined meta-analysis for stages 1, 2 and 3. We combined stage 1, 2 and 3 data using both meta-analysis approaches (fixed-effects model to combine ORs and weighted P value–based z statistic combination across all sample sets) described above. We also assessed our results using random effects meta-analysis (**Supplementary Table 6**). We observed some evidence for heterogeneity across studies (the I^2 statistic ranged from 0 to 57.8% depending on the SNP), with rs7578597 and rs10923931 showing the largest fold differences in association P value between the fixed- and random-effects model analyses. Differences in strength of association across studies (leading to evidence for heterogeneity) could reflect interesting biological associations that vary from study to study depending on subject ascertainment scheme.

Genomic control. An expanded description of these methods is described in **Supplementary Methods**. We adopted two strategies in reporting the findings from this study. In the first, we performed GC-correction of data from DGI, FUSION and WTCCC before stage 1 meta-analysis. We corrected each individual study for the GC inflation observed (directly genotyped and imputed data separately), and combined results across studies. We present the genome-wide distribution of association statistics in **Supplementary Figure 1**. We note that, after study-specific genomic control adjustment, the estimated inflation factor for the stage 1 meta-analysis test statistic was 1.04.

In the second strategy, we combined GC-uncorrected data from DGI, FUSION and WTCCC for stage 1 meta-analysis and did not correct the meta-analysis test statistics for the overall GC (to guard against over-conservativeness in the estimate of strength of association for interesting signals). We also present the genome-wide distribution of these statistics in **Supplementary Figure 1**.

For the combination of data across stages 1, 2 and 3, we also adopted these two strategies (of using GC-corrected and GC-uncorrected stage 1 data). In the first, we performed individual GC-correction of DGI, FUSION and WTCCC stage 1 data before meta-analysis with stage 2 and stage 3 data (an approach which may be over-conservative where, as was the case here, none of the T2D-associated SNPs had particular hallmarks of stratification) (**Supplementary Note**). In the second, we combined only uncorrected data (except for the deCODE data, for which we applied GC correction, given a more marked genomic control inflation (GC $\sim$1.3) in that sample). We present the resulting data from both approaches (of using GC-corrected and GC-uncorrected stage 1 data for stage 1–3 meta-analysis) in **Supplementary Table 6** and a comparison of results (showing very small differences) in the **Supplementary Note**. All data presented elsewhere in the manuscript reflect the GC-corrected analysis strategy outcome.

Conditional analysis of T2D signals. For each SNP in **Table 2**, we assessed the additive SNP association in the stage 1 and 2 samples before and after including

BMI in the logistic regression model. For each genotyped and imputed SNP surrounding a specific T2D signal, we assessed the additive SNP association in the stage 1 sample before and after including the **Table 2** SNP from the same region in the model. We analyzed the data and adjusted for covariates for the stage 1 and stage 2 analysis of each sample. Data were combined across studies as described above. The ORs and CIs were calculated using a fixed-effects model, and *P* values were calculated using the weighted *z* score method. For the WTCCC stage 1 samples, we did not have BMI information available for ~1,500 of the population-based controls. We therefore carried out the conditional BMI analyses by using all T2D cases and only those controls for whom BMI data were available.

Quantitative trait analyses. Quantitative trait analyses were carried out in the UK, DGI and FUSION samples for the 11 SNPs taken forward to stage 3. We tested BMI, quantitative glycemic traits (fasting and 2-h levels of glucose and insulin, HOMA-IR (homeostasis model assessment of insulin resistance)), lipid traits (total, HDL and LDL cholesterol, and serum triglycerides) and blood pressure (systolic and diastolic), where available, for association using an additive genetic model (**Supplementary Methods**).

URLs. MACH, http://www.sph.umich.edu/csg/abecasis/MaCH/download.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

UK: Collection of the UK type 2 diabetes cases was supported by Diabetes UK, BDA Research and the UK Medical Research Council (Biomedical Collections Strategic Grant G0000649). The UK Type 2 Diabetes Genetics Consortium collection was supported by the Wellcome Trust (Biomedical Collections Grant GR072960). The GWA genotyping was supported by the Wellcome Trust (076113), and the replication genotyping was supported by the European Commission (EURODIA LSHG-CT-2004- 518153), MRC (Project Grant G0601261), Wellcome Trust, Peninsula Medical School and Diabetes UK. E.Z. is a Wellcome Trust Research Career Development Fellow. We acknowledge the contribution of M. Sampson and our team of research nurses. We acknowledge the efforts of J. Collier, P. Robinson, S. Asquith and others at KBioscience for their rapid and accurate large-scale genotyping.

DGI: We thank the study participants who made this research possible. We thank colleagues in the Broad Genetic Analysis and Biological Samples Platforms for their expertise and contributions to genotyping, data and sample management, and analysis. The initial GWAS genotyping was supported by Novartis (to D.A.); support for additional analysis and genotyping in this report was provided by funding from the Broad Institute of Harvard and MIT, by the Richard and Susan Smith Family Foundation/American Diabetes Association Pinnacle Program Project Award (to D.A.), and by a Freedom to Discovery award of the Foundation of Bristol Myers Squibb (to D.A.). P.I.W.dB., M.J.D. and D.A. acknowledge support from US National Institutes of Health/National Heart, Lung, and Blood Institute grant (U01 HG004171). D.A. was a Burroughs Wellcome Fund Clinical Scholar in Translational Research and is a Distinguished Clinical Scholar of the Doris Duke Charitable Foundation. L.G., T.T., B.I. and M.R.T. and the Botnia Study are principally supported by the Sigrid Juselius Foundation, the Finnish Diabetes Research Foundation, The Folkhalsan Research Foundation and Clinical Research Institute HUCH Ltd; work in Malmö, Sweden was also funded by a Linné grant from the Swedish Research Council (349-2006-237). We thank the Botnia and Skara research teams for clinical contributions, and colleagues at MGH, Harvard, Broad, Novartis and Lund for helpful discussions throughout the course of this work.

FUSION: We thank the Finnish citizens who generously participated in this study and R.Welch for bioinformatics support. Support for this research was provided by US National Institutes of Health grants DK062370 (M.B.), DK072193 (K.L.M.), HL084729 (G.R.A.), HG002651 (G.R.A.) and U54 DA021519; National Human Genome Research Institute intramural project number 1 Z01 HG000024 (F.S.C.); and a postdoctoral fellowship award from the American Diabetes Association (C.J.W.). Genome-wide genotyping was performed by the Johns Hopkins University Genetic Resources Core Facility (GRCF) SNP Center at the Center for Inherited Disease Research (CIDR) with support from CIDR NIH Contract Number N01-HG-65403 and the GRCF SNP Center.

deCODE: We thank the Icelandic study participants whose contribution made this work possible. We also thank the nurses at Noatun (deCODE's sample recruitment center) and personnel at the deCODE core facilities.

KORA study: We thank C. Gieger and G. Fischer for expert data handling. The MONICA/KORA Augsburg studies were financed by the GSF-National Research Center for Environment and Health, Neuherberg, Germany and supported by grants from the German Federal Ministry of Education and Research (BMBF). Part of this work was financed by the German National Genome Research Network (NGFN). Our research was also supported within the Munich Center of Health Sciences (MC Health) as part of LMUinnovativ. We thank all members of field staffs who were involved in the planning and conduct of the MONICA/ KORA Augsburg studies.

Danish study: This work was supported by the European Union (EUGENE2, grant no. LSHM-CT-2004-512013), Lundbeck Foundation centre of Applied Medical Genomics in Personalized Disease Prediction, Prevention and Care and The Danish Medical Research Council.

HUNT: The Nord-Trøndelag Health Study (The HUNT Study) is a collaboration between The HUNT Research Centre, Faculty of Medicine, Norwegian University of Science and Technology (NTNU), The National Institute of Public Health, The National Screening Service of Norway and The Nord-Trøndelag County Council.

NHS: The Nurses' Health Study is funded by National Cancer Institute grant CA87969. L.Q. is supported by an American Heart Association Scientist Development Grant. F.B.H. is supported by NIH grants DK58845 and U01 HG004399.

GEM Consortium: We thank all study participants. The work on the Cambridgeshire case-control, Ely, ADDITION and EPIC-Norfolk studies was funded by support from the Wellcome Trust and MRC. The Norfolk Diabetes study is funded by the MRC with support from NHS Research & Development and the Wellcome Trust. We are grateful to S. Griffin, MRC Epidemiology Unit, for assistance with the ADDITION study and M. Sampson and E. Young for help with the Norfolk Diabetes Study. We thank S. Bumpstead, W.E. Bottomley and A. Chaney for rapid and accurate genotyping and J. Ghori for assay design and informatics support. We are grateful to P. Deloukas for overall genotyping support. F.P. and I.B. are funded by the Wellcome Trust.

METSIM: The METSIM study has received grant support from the Academy of Finland (no. 124243).

AUTHOR CONTRIBUTIONS

Writing team and project management: L.J.S., E.Z., R.S., B.F.V., D.A., M.B. & M.I.M. **Study design:** R.S., B.F.V., E.Z., L.J.S., T.E.H., F.B.H., J.J.R., H.C., K.S., O.P., T.L., K.H., M.L., A.T.H., I.B., N.J.W., F.S.C., L.G., D.A., M.I.M. & M.B. **Analysis:** K.S.E., R.M.F., H.L., C.M.L., J.R.B.P., I.P., N.W.R., N.J.T., M.N.W., J.L.M. & E.Z. (UK), P.S.C., C.-J.D., W.L.D., T. Hu, A.U.J., Y.L., H.M.S., C.J.W., G.R.A. & L.J.S. (FUSION), R.S., B.F.V., P.I.W.dB., F.G.K., P.A. & M.J.D. (DGI), U.T. & A.K. (deCODE), N.G., G.A., T.H. & O.P. (Danish), K.M. (HUNT), L. Qi (NHS), C.L. (GEM Consortium), M.L. (Metsim). **Clinical samples and genotyping:** WTCCC, A.S.F.D., T.M.F., C.J.G., G.A.H., K.R.O., C.N.A.P., B.S., M.W., A.D.M., A.T.H. & M.I.M. (UK), L.L.B., P.D., M.R.E., K.K., M.A.M., N.N., M.R., A.J.S., N.N.B., K.L.M., J.T., A.F.M., L. Qin & R.M.W. (FUSION), K.A., K.B.B., N.P.B., L. Gianniny, C.G., B.I., V.L., P.N., M.S., T.T. & L. Groop (DGI), V.S., G.T. & K.S. (deCODE), H.G., C.H., C.M. & T.I. (KORA), G.A., N.G., T.H., T.J., T.L., A.S., K.B.-J. & O.P. (Danish), K.M., E.P., C.P. & K.H. (HUNT), F.B.H. (NHS), F.P., I.B. & N.J.W. (GEM Consortium), J.K. (METSIM).

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturegenetics/.

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Diabetes Genetics Initiative.. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316**, 1331–1336 (2007).
2. Scott, L.J. *et al.* A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* **316**, 1341–1345 (2007).
3. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
4. Zeggini, E. *et al.* Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* **316**, 1336–1341 (2007).
5. Steinthorsdottir, V. *et al.* A variant in *CDKAL1* influences insulin response and risk of type 2 diabetes. *Nat. Genet.* **39**, 770–775 (2007).

6. Sladek, R. *et al.* A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881–885 (2007).
7. Florez, J.C. *et al.* A 100K genome-wide association scan for diabetes and related traits in the Framingham Heart Study replication and integration with other genome-wide datasets. *Diabetes* **56**, 3063–3074 (2007).
8. Rampersaud, E. *et al.* Identification of novel candidate genes for type 2 diabetes from a genome-wide association scan in the Old Order Amish: evidence for replication from diabetes-related quantitative traits and from independent populations. *Diabetes* **56**, 3053–3062 (2007).
9. Hanson, R.L. *et al.* A search for variants associated with young-onset type 2 diabetes in American Indians in a 100K genotyping array. *Diabetes* **56**, 3045–3052 (2007).
10. Hayes, M.G. *et al.* Identification of type 2 diabetes genes in Mexican Americans through genome-wide association studies. *Diabetes* **56**, 3033–3044 (2007).
11. Salonen, J. *et al.* Type 2 diabetes whole-genome association study in four populations: the DiaGen consortium. *Am. J. Hum. Genet.* **81**, 338–345 (2007).
12. McCarthy, M.I. & Zeggini, E. Genome-wide association scans for Type 2 diabetes: new insights into biology and therapy. *Trends Pharmacol. Sci.* **28**, 598–601 (2007).
13. International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
14. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906–913 (2007).
15. Freedman, M.L. *et al.* Assessing the impact of population stratification on genetic association studies. *Nat. Genet.* **36**, 388–393 (2004).
16. Nakajima, T., Fujino, S., Nakanishi, G., Kim, Y.S. & Jetten, A.M. TIP27: a novel repressor of the nuclear orphan receptor TAK1/TR4. *Nucleic Acids Res.* **32**, 4194–4204 (2004).
17. Collins, L.L. *et al.* Growth retardation and abnormal maternal behavior in mice lacking testicular orphan nuclear receptor 4. *Proc. Natl. Acad. Sci. USA* **101**, 15058–15063 (2004).
18. Thomas, G. *et al.* Multiple loci identified in a genome-wide association study of prostate cancer. *Nat. Genet.* **40**, 310–315 (2008).
19. Gudmundsson, J. *et al.* Two variants on chromosome 17 confer prostate cancer risk, and the one in *TCF2* protects against type 2 diabetes. *Nat. Genet.* **39**, 977–983 (2007).
20. Winckler, W. *et al.* Evaluation of common variants in the six known maturity-onset diabetes of the young (MODY) genes for association with type 2 diabetes. *Diabetes* **56**, 685–693 (2007).
21. Bieganowski, P., Shilinski, K., Tschlis, P.N. & Brenner, C. Cdc123 and checkpoint forkhead associated with RING proteins control the cell cycle by controlling eIF2γ abundance. *J. Biol. Chem.* **273**, 44656–44666 (2004).
22. Driescher, N. *et al.* Evidence for a 3p25 breakpoint hot spot region in thyroid tumors of follicular origin. *Thyroid* **16**, 1091–1096 (2006).
23. Driescher, N. *et al.* A domain of the thyroid adenoma associated gene (THADA) conserved in vertebrates becomes destroyed by chromosomal rearrangements observed in thyroid adenomas. *Gene* **403**, 110–117 (2007).
24. Lammert, E., Brown, J. & Melton, D.A. Notch gene expression during pancreatic organogenesis. *Mech. Dev.* **94**, 199–203 (2000).
25. Higgins, J.P., Thompson, S.G., Deeks, J.J. & Altman, D.G. Measuring inconsistency in meta-analyses. *Br. Med. J.* **327**, 557–560 (2003).
26. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
27. Lin, S., Chakravarti, A. & Cutler, D.J. Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies. *Nat. Genet.* **36**, 1181–1188 (2004).

The complete list of authors is as follows:

Elefheria Zeggini[1,10], Laura J Scott[2,10], Richa Saxena[3–8,10], Benjamin F Voight[3–5,7,10], Jonathan L Marchini[11], Tianle Hu[2], Paul IW de Bakker[3,7,12], Gonçalo R Abecasis[2], Peter Almgren[13], Gitte Andersen[14], Kristin Ardlie[3], Kristina Bengtsson Boström[15], Richard N Bergman[16], Lori L Bonnycastle[17], Knut Borch-Johnsen[14,18], Noël P Burtt[3], Hong Chen[19], Peter S Chines[17], Mark J Daly[3–5,7], Parimal Deodhar[17], Chia-Jen Ding[2], Alex S F Doney[20], William L Duren[2], Katherine S Elliott[1], Michael R Erdos[17], Timothy M Frayling[21,22], Rachel M Freathy[21,22], Lauren Gianniny[3], Harald Grallert[23], Niels Grarup[14], Christopher J Groves[24], Candace Guiducci[3], Torben Hansen[14], Christian Herder[25], Graham A Hitman[26], Thomas E Hughes[19], Bo Isomaa[27,28], Anne U Jackson[3], Torben Jørgensen[29], Augustine Kong[30], Kari Kubalanza[17], Finny G Kuruvilla[3,4,6], Johanna Kuusisto[31], Claudia Langenberg[32], Hana Lango[21,22], Torsten Lauritzen[33], Yun Li[2], Cecilia M Lindgren[1,24], Valeriya Lyssenko[13], Amanda F Marvelle[34], Christa Meisinger[23], Kristian Midthjell[35], Karen L Mohlke[34], Mario A Morken[17], Andrew D Morris[20], Narisu Narisu[17], Peter Nilsson[15], Katharine R Owen[24], Colin NA Palmer[36], Felicity Payne[37], John R B Perry[21,22], Elin Pettersen[38], Carl Platou[39], Inga Prokopenko[1,24], Lu Qi[39,40], Li Qin[34], Nigel W Rayner[1,24], Matthew Rees[37], Jeffrey J Roix[19], Anelli Sandbæk[18], Beverley Shields[22], Marketa Sjögren[13], Valgerdur Steinthorsdottir[30], Heather M Stringham[2], Amy J Swift[17], Gudmar Thorleifsson[30], Unnur Thorsteinsdottir[30], Nicholas J Timpson[1,41], Tiinamaija Tuomi[28,42], Jaakko Tuomilehto[43–45], Mark Walker[46], Richard M Watanabe[47], Michael N Weedon[21,22], Cristen J Willer[2], Wellcome Trust Case Control Consortium[49], Thomas Illig[23], Kristian Hveem[35], Frank B Hu[39,40], Markku Laakso[31], Kari Stefansson[30], Oluf Pedersen[14,18], Nicholas J Wareham[32], Inês Barroso[37], Andrew T Hattersley[21,22], Francis S Collins[17], Leif Groop[13,42], Mark I McCarthy[1,24,50], Michael Boehnke[2,50] & David Altshuler[3,4,6–8,48,50]

[11]Department of Statistics, University of Oxford, Oxford, OX1 3TG, UK. [12]Division of Genetics, Brigham and Women's Hospital, Harvard-Partners Center for Genetics and Genomics, Boston, Massachusetts 02115, USA. [13]Department of Clinical Sciences, Diabetes and Endocrinology Research Unit, University Hospital Malmö, Lund University, S-205 02 Malmö, Sweden. [14]Steno Diabetes Center, Copenhagen, DK-2820, Denmark. [15]Skaraborg Institute, S-541 30 Skövde, Sweden. [16]Department of Physiology and Biophysics, Keck School of Medicine, University of Southern California, Los Angeles, California 90033, USA. [17]Genome Technology Branch, National Human Genome Research Institute, Bethesda, Maryland 20892, USA. [18]Faculty of Health Science, University of Aarhus, Aarhus, DK-8000, Denmark. [19]Diabetes and Metabolism Disease Area, Novartis Institutes for BioMedical Research, 1DO Technology Square, Cambridge, Massachusetts 02139, USA. [20]Diabetes Research Group, Division of Medicine and Therapeutics, Ninewells Hospital and Medical School, Dundee, DD1 9SY, UK. [21]Genetics of Complex Traits, Institute of Biomedical and Clinical Science, Peninsula Medical School, Magdalen Road, Exeter, EX1 2LU, UK. [22]Diabetes Genetics, Institute of Biomedical and Clinical Science, Peninsula Medical School, Barrack Road, Exeter, EX2 5DW, UK. [23]Gesellschaft für Strahlenforschung-National Research Center for Environment and Health, Institute of Epidemiology, D-85764 Neuherberg, Germany. [24]Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Oxford, OX3 7LJ, UK. [25]Institute for Clinical Diabetes Research, German Diabetes Center, Leibniz Institute at Heinrich Heine University, D-40225 Düsseldorf, Germany. [26]Centre for Diabetes and Metabolic Medicine, Barts and The London, Royal London Hospital, Whitechapel, London, E1 1BB, UK. [27]Malmska Municipal Health Center and Hospital, FIN-68601 Jakobstad, Finland. [28]Folkhälsan Research Center, FIN-00D14 Helsinki, Finland. [29]Research Centre for Prevention and Health, Glostrup University Hospital, DK-2600 Glostrup, Denmark. [30]deCODE genetics, Sturlugata 8, IS-101 Reykjavik, Iceland. [31]Department of Medicine, University of Kuopio and Kuopio University Hospital, 70210, Kuopio, Finland. [32]MRC Epidemiology Unit, Institute of Metabolic Science, Addenbrooke's Hospital, Cambridge CB2 0QQ, UK. [33]Department of General Practice, University of Aarhus, DK-800D Aarhus, Denmark. [34]Department of Genetics, University of North Carolina, Chapel Hill, North Carolina 27599, USA. [35]HUNT Research Centre, Department of Public Health and General Practice, Faculty of Medicine, Norwegian University of Science and Technology (NTNU), 7650 Verdal, Norway. [36]Population Pharmacogenetics Group, Biomedical Research Centre, Ninewells Hospital and Medical School, Dundee, DD1 9SY, UK. [37]Metabolic Disease Group, Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1SA, UK. [38]Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology (NTNU), N-7600 Levanger, Norway. [39]Departments of Nutrition and Epidemiology, Harvard School of Public Health, Boston, Massachusetts 02115, USA. [40]Channing Laboratory, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts D2115, USA. [41]The MRC Centre for Causal Analyses in Translational Epidemiology, Bristol University, Canynge Hall, Whitelades Road, Bristol, BS2 8PR, UK. [42]Department of Medicine, Helsinki University Hospital, University of Helsinki, FIN-0D300 Helsinki, Finland. [43]Diabetes Unit, Department of Epidemiology and Health Promotion, National Public Health Institute, 00300 Helsinki, Finland. [44]Department of Public Health, University of Helsinki, D0014 Helsinki, Finland. [45]South Dstrobothnia Central Hospital, 60220 Seinäjoki, Finland. [46]Diabetes Research Group, School of Clinical Medical Sciences, Newcastle University, Framlington Place, Newcastle upon Tyne, NE2 4HH, UK. [47]Department of Preventative Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California 90033, USA. [48]Diabetes Unit, Massachusetts General Hospital, Boston, Massachusetts D2114, USA. [49]Membership of the Wellcome Trust Case Control Consortium is provided in the Supplementary Note. [50]These authors contributed equally to this work.